UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

UFC
UNIVERSITÉ
DE FRANCHE-COMTÉ

THÈSE DE DOCTORAT

DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE FRANCHE-COMTÉ

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Informatique

par

HÉBER HWANG ARCOLEZI

Production of Categorical Data Verifying Differential Privacy: Conception and Applications to Machine Learning

Production de Données Catégorielles Respectant la Confidentialité Différentielle : Conception et Applications au Apprentissage Automatique

Thèse présentée et soutenue à Besançon, le 05 Janvier 2022

Composition du Jury :

PROF CHRÉTIEN STÉPHANE	Université de Lyon 2	Président
PROF CUNCHE MATHIEU	Institut National des Sciences Appliquées de Lyon	Rapporteur
PROF NGUYEN BENJAMIN	Institut National des Sciences Appliquées Centre Val de Loire	Rapporteur
PROF ALVIM MÁRIO S.	Universidade Federal de Minas Gerais	Examinateur
PROF COUCHOT JEAN-FRANÇOIS	Université Bourgogne Franche-Comté	Directeur de thèse
PROF XIAO XIAOKUI	National University of Singapore	Codirecteur de thèse

N° | X | X | X |

ABSTRACT

Production of Categorical Data Verifying Differential Privacy: Conception and Applications to Machine Learning

Héber Hwang Arcolezi
University Bourgogne Franche Comté, 2022

Supervisors: Jean-François Couchot, Bechara Al Bouna, and Xiaokui Xiao

Private and public organizations regularly collect and analyze digitalized data about their associates, volunteers, clients, etc. However, because most personal data are sensitive, there is a key challenge in designing privacy-preserving systems to comply with data privacy laws, e.g., the General Data Protection Regulation. To tackle privacy concerns, research communities have proposed different methods to preserve privacy, with Differential privacy (DP) standing out as a formal definition that allows quantifying the privacy-utility trade-off. Besides, with the local DP (LDP) model, users can sanitize their data locally before transmitting it to the server.

The objective of this thesis is thus two-fold: $\mathbf{O}_1$) To improve the utility and privacy in multiple frequency estimates under LDP guarantees, which is fundamental to *statistical learning*. And $\mathbf{O}_2$) To assess the privacy-utility trade-off of machine learning (ML) models trained over differentially private data.

For $\mathbf{O}_1$, we first tackled the problem from two "*multiple*" perspectives, i.e., multiple attributes and multiple collections throughout time (longitudinal studies), while focusing on **utility**. Secondly, we focused our attention on the multiple attributes aspect only, in which we proposed a solution focusing on **privacy** while preserving utility. In both cases, we demonstrate through analytical and experimental validations the advantages of our proposed solutions over state-of-the-art LDP protocols.

For $\mathbf{O}_2$, we empirically evaluated ML-based solutions designed to solve real-world problems while ensuring DP guarantees. Indeed, we mainly used the *input data perturbation* setting from the privacy-preserving ML literature. This is the situation in which the whole

dataset is *sanitized* independently (i.e., row-by-row) and, thus, we implemented LDP algorithms from the perspective of the centralized data owner. In all cases, we concluded that differentially private ML models achieve nearly the same utility metrics as non-private ones.

KEYWORDS: Differential privacy, Local differential privacy, Categorical data, Machine learning.

RÉSUMÉ

Production de Données Catégorielles Respectant la Confidentialité Différentielle : Conception et Applications au Apprentissage Automatique

Héber Hwang Arcolezi
Université Bourgogne Franche Comté, 2022

Encadrants: Jean-François Couchot, Bechara Al Bouna, et Xiaokui Xiao

Les organisations privées et publiques collectent et analysent régulièrement des données numérisées sur leurs associés, volontaires, clients, etc. Cependant, comme la plupart des données personnelles sont sensibles, la conception de systèmes préservant la vie privée pour se conformer aux lois sur la confidentialité des données, par exemple le règlement général sur la protection des données, constitue un défi important. Pour résoudre les problèmes de confidentialité, les communautés de chercheurs ont proposé différentes méthodes de préservation de la confidentialité, la confidentialité différentielle (DP) se distinguant comme une définition formelle qui permet de quantifier le compromis entre confidentialité et utilité. En outre, avec le modèle de confidentialité différentielle locale (LDP), les utilisateurs peuvent sanitisé leurs données localement avant de les transmettre au serveur.

L'objectif de cette thèse est donc double : $\mathbf{O}_1$) Améliorer l'utilité et la confidentialité des estimations de fréquences multiples sous garanties LDP, ce qui est fondamental pour *l'apprentissage statistique*. Et $\mathbf{O}_2$) Évaluer le compromis vie privée-utilité des modèles d'apprentissage machine (ML) entraînés sur des données différentiellement privées.

Pour $\mathbf{O}_1$, nous avons premièrement abordé le problème sous deux angles "*multiple*", c'est-à-dire des attributs multiples et des collections multiples dans le temps (études longitudinales), tout en nous concentrant sur **utilité**. Deuxièmement, nous avons concentré notre attention sur l'aspect des attributs multiples uniquement, dans lequel nous avons proposé une solution axée sur la **confidentialité** tout en préservant l'utilité. Dans les deux cas, nous démontrons par des validations analytiques et expérimentales les avan-

tages de nos solutions proposées par rapport aux protocoles LDP de pointe.

Pour $\mathbf{O}_2$, nous avons évalué empiriquement des solutions basées sur les ML conçues pour résoudre des problèmes du monde réel tout en assurant des garanties de DP. En effet, nous avons principalement utilisé le cadre *perturbation des données d'entrée* de la littérature sur les ML préservant la confidentialité. Il s'agit de la situation dans laquelle l'ensemble des données est *sanitisé* indépendamment (c'est-à-dire ligne par ligne) et, par conséquent, nous avons mis en œuvre des algorithmes LDP du point de vue du propriétaire centralisé des données. Dans tous les cas, nous avons conclu que les modèles ML différentiellement privés atteignent presque les mêmes mesures d'utilité que les modèles non privés.

Mots clés: Confidentialité différentielle, Confidentialité différentielle locale, Données catégorielles, Apprentissage automatique.

ACKNOWLEDGEMENTS

Primarily, I would like to express my greatest thanks to my supervisor, Professor Jean-François Couchot, for his support, leadership, and encouragement during my Ph.D. study. I am very fortunate to have had him as my supervisor and for being led toward a topic I am very passionate about. I am truly grateful for his personality as an advisor as Jean-François really cares about his students, in both academic and personal subjects, which I wish for any Ph.D. student to have. I also thank my co-supervisors Bechara Al Bouna and Xiaokui Xiao for their collaboration and support throughout this dissertation.

I would also like to thank Professors Benjamin Nguyen, Mathieu Cunche, Stéphane Chrétien, and Mário S. Alvim, who kindly accepted to be part of my dissertation jury and for their valuable suggestions on research perspectives.

Thanks also to Denis Renaud, who leads the Orange Application for Business team in Belfort, for his continued collaboration and helpful feedback. I also thank Commandant Guillaume Royer-Fey and Capitaine Céline Chevallier from the Fire Department of Doubs and Professor Christophe Guyeux, who helped me a lot through fruitful collaboration and a lot of feedback.

I also thank Professor Sébastien Gambs, who kindly mentored me during my research visit at the Université du Québec à Montréal, and for the opportunity to continue collaborating. I learned a lot from him and gained valuable experiences, which are important for my career as a researcher.

I am very, very grateful to Selene Cerna, a special person to me, for the many joyful moments, constant support, and for taking care of me all these years. Selene has supported me since my master's degree and was significant in my growth as a young researcher. I admire Selene for her great willingness to help and share with others, and I am fortunate to be one of those people. I learned a lot with her, both technically and through extensive discussion on research subjects, which essentially helped me during this Ph.D. study.

I also thank Zhì Háo Chen who gave me a lot of guidance through many bureaucratic processes to establish me as a foreign doctoral student in France.

Last but not least, my beloved grandparents, parents, and siblings, my biggest thank to each of you who have supported and cared for me throughout my life. From each of you, a different kind of love has been shown over the years, and I gladly consider and return all the love I can offer you all.

CONTENTS

List of abbreviations

ACC . . . . . Accuracy

ADP . . . . . Adaptive

CDRs . . . . . Call Detail Records

CNIL . . . . . Commission Nationale de l'Informatique et des Libertés

COVID-19 . . . Coronavirus Disease 2019

DP . . . . . Differential privacy

EMS . . . . . Emergency medical services

FIMU . . . . . Festival International de Musiques Universitaires

GDPR . . . . . General Data Protection Regulation

GRR . . . . . Generalized Randomized Response

LDP . . . . . Local Differential Privacy

LP . . . . . Linear Program

MF1 . . . . . Macro F1-Score

ML . . . . . Machine Learning

MNO . . . . . Mobile Network Operator

MSE . . . . . Mean Squared Error

MS-FIMU . . . . Mobility Scenario FIMU

OBS . . . . . Orange Business Services

OUE . . . . . Optimized Unary Encoding

QID . . . . . Quasi-Identifier

RMSE . . . . . Root Mean Square Error

RR . . . . . Randomized Response

Smp . . . . . Sampling

Spl . . . . . Splitting

SUE . . . . . Symmetric Unary Encoding

UE . . . . . Unary Encoding

I

THESIS INTRODUCTION

1

INTRODUCTION

1.1/ INTRODUCTION

Let be given Article 12 from the Universal Declaration of Humans Right [10], which defines: "*No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks*."

Notice, however, that with the advancement of technology of information (**not only correspondences anymore**), protecting individuals' privacy in the era of Big data is a significant challenge. Indeed, the explosion of the number of connected objects, mobile applications collecting and/or generating any type of data makes personal data ubiquitous and growing exponentially.

Moreover, when collecting data in practice, one is often interested in multiple attributes of a population, i.e., *multidimensional data*. For instance, in crowd-sourcing applications, the server may collect both demographic information (e.g., gender, nationality) and user habits in order to develop personalized solutions for specific groups. In addition, one generally aims to collect data from the same users throughout time (i.e., *longitudinal* studies), which is essential in many situations. For example, the fact that remote antennas of mobile network operators (MNOs) have received cell phone connections may reveal a movement if the same user is identified in different antennas throughout time.

From a human point of view, data analysts can be external providers. In other words, they very rarely have the consent of the data providers (i.e., individuals concerned) to analyze the data. It is, therefore, necessary for the company providing the service to make all possible efforts to follow all the recommendations from data privacy authorities such as the General Data Protection Regulation (GPDR) [112] and, particularly, make any re-identification unfeasible from a practical point of view. On the other hand, even if trusted service providers collect raw personal data, this practice can still lead to privacy breaches, i.e., the risk of information leakage is always possible even if service providers

make every effort to secure the data.

Indeed, data breaches are all too common [228], which endanger users' privacy and can lead to substantial losses for companies under the GDPR (cf. [126, 167], for example). Moreover, along with gathering data, extracting high-utility analytics through machine learning (ML) from the collected data is of great interest. Yet, even ML models trained with raw data can also indirectly reveal sensitive information [185, 54] (e.g., cf. [105, 104, 145]).

In addition, privacy issues appear more than ever in headlines (e.g., [64, 24, 97, 58, 236, 223, 227]). To tackle privacy concerns, research communities have proposed **different methods to preserve privacy**, in which the **main goal is that anonymized data should not leak private information about any individual** [181]. To this end, k-anonymity [18, 20] and differential privacy (DP) [27, 26, 59] are two well-known privacy techniques. On the one hand, k-anonymity is very risky since it does not allow to counter intersecting and/or homogeneity attacks, for example [28, 29]. On the other hand, DP has been increasingly accepted as the current standard for data privacy [73, 132, 220, 59]. However, in the originally proposed centralized DP model, queries perturbed by DP algorithms require the storage of raw databases because the noise is only added at the end of the request. As aforementioned, storing and/or sharing raw databases (as well as training ML models over raw data) is not always desirable because it is necessary to secure all access to them from both a technical and human point of view.

To preserve privacy at the user-side, an alternative approach, namely, local differential privacy (LDP), was initially formalized in [32]. With LDP, rather than trusting in a data curator to have the raw data and sanitize it to output queries, each user applies a DP mechanism to their data before transmitting it to the data collector server. The LDP model allows collecting data in unprecedented ways and, therefore, has led to several adoptions by industry. For instance, big tech companies like Google, Apple, and Microsoft, reported the implementation of LDP mechanisms to gather statistics in well-known systems (i.e., Google Chrome browser [61], Apple iOS and macOS [106], and Windows 10 operation system [95]).

1.2/ Motivation and Objectives

For the rest of this manuscript, the author will utilize **we** rather than **I** to highlight the contributions of all my collaborators (cf. Acknowledgment on page vii). Yet, **the author is the only one responsible for all errors** that may still be present on this manuscript. The work in this manuscript is based on two motivating projects.

On the one hand, we had a preliminary collaboration with the Orange Business Services

(OBS) team in Belfort, France, i.e., an MNO. The OBS team presented us *an overview* of their deployed system named Flux Vision [53], which publishes real-time statistics on human mobility by analyzing call detail records (CDRs). The Flux Vision system motivated us **to study how to gather knowledge from the published statistics as well as to propose a distinct privacy-preserving data collection process**. More precisely, from a practical perspective, based on *longitudinal* and *multidimensional* OBS mobility reports, we noticed that these statistics could be improved to provide more information about mobility patterns of the individuals concerned. Thus, this is our **first objective**. Furthermore, our **second objective** is to propose a privacy-preserving CDRs processing system, which could improve the privacy of MNOs' clients. Next, from a theoretical perspective on statistical learning, our **third objective** is to improve the utility and privacy of multiple frequency estimates (i.e., multidimensional and longitudinal data collections) under LDP guarantees.

In addition, we also worked on a collaborative framework with Selene Cerna and Christophe Guyeux, members of the AND[1] research team from the same research department as ours[2]. Selene Cerna holds a CIFRE thesis (N 2019/0372) with the fire department named Service Départemental d'Incendie et de Secours du Doubs (SDIS 25), i.e, an emergency medical services (EMS) in France. For the past few years, the AND team has been investigating ML-based solutions to optimize the SDIS 25 services under a strict confidentiality agreement on the SDIS 25 data. The way these data have been shared motivated us **to study the privacy-utility trade-off of ML models trained over sanitized data**. That is, we consider the case of centralized data owners (e.g., MNOs and EMS) that *collect sensitive information* from individuals for both billing and/or legal purposes *but do not trust* the third entity to develop decision-support systems. So, our **fourth and last objective** is to evaluate empirically the privacy-utility trade-off of different ML-based solutions trained over sanitized data. We mainly focused on the SDIS 25 data. Notice, however, that this manuscript *does not* focus on the data collection nor the feature engineering processes carried out by Selene Cerna but, rather, we will present only necessary information about the dataset while *focusing on the privacy-utility trade-off* analysis.

1.3/ Main Contributions of this Thesis

The main contributions of this thesis are summarized in the following:

1. First, based on one-week statistical data of unions of consecutive days published by OBS [53], we present a method for inferring and recreating a synthetic dataset that

[1] Algorithmique Numérique Distribuée (or, distributed digital algorithmics in English).
[2] Department of Informatics and Complex Systems (DISC in French).

matches the original statistical data with low mean relative error. We thus generated and published it as an open dataset (https://github.com/hharcolezi/OpenMSFIMU) such that others can use it to evaluate new privacy-preserving techniques as well as ML tasks.

2. Second, by studying these aggregate statistics on human mobility, we proposed an LDP-based CDRs processing system to generate multidimensional mobility reports throughout time by offering strong privacy guarantees for each user.

3. The first two studies on CDRs-based mobility reports are translated to longitudinal statistical releases about the frequency of visitors by multiple attributes. We then contribute to the **theoretical** aspect under the LDP setting. More precisely, we first focused on optimizing the *utility* of LDP protocols for *longitudinal* and *multidimensional* frequency estimates.

4. Next, we identified a limitation of the state-of-the-art solution used for multidimensional frequency estimates with LDP, which splits users into groups instead of splitting the privacy budget. We then propose a solution to this limitation, which improves the *privacy* of users while providing *the same or better utility* (regarding the mean squared error metric) than the state-of-the-art solution.

5. Lastly, we empirically evaluated the privacy-utility trade-off of differentially private input perturbation-based ML models. That is, we assessed practical solutions in which data owners (e.g., MNOs and EMS) could sanitize their datasets locally before transmitting these data to untrusted parties to develop decision-support tools, with no considerable impact on the utility.

1.4/ THESIS OUTLINE

The rest of this manuscript is organized as follows: Chapter 2 presents the scientific background on data anonymization techniques. Chapter 3 provides the scientific background on machine learning techniques and presents the databases we will experiment on. Chapter 4 presents the first contribution of this manuscript, namely, an open, longitudinal, and synthetic dataset of faked virtual humans generated by an optimization approach applied to a real-life CDRs-based anonymized database. Chapter 5 proposes a privacy-preserving CDRs processing system to generate mobility reports longitudinally. Chapter 6 presents our first theoretical contribution on statistical learning with LDP. Chapter 7 resolves one limitation of Chapters 5 and 6 by improving the privacy of individuals while keeping the utility on statistical learning with LDP. Chapter 8 empirically evaluates two differentially private machine learning settings on multivariate time series forecasting. Chapter 9 proposes a privacy-preserving methodology to sanitize an EMS intervention

dataset while allowing both statistical learning and forecasting tasks. Chapter 10 empirically evaluates the impact of sanitizing the location of an emergency when training ML models to predict the response time of ambulances. Chapter 11 empirically evaluates the impact of training ML models over anonymized data to predict the victims' mortality. Lastly, Chapter 12 provides a general conclusion of this work and its perspectives.

II

BACKGROUND

2

Data Anonymization

In Chapter 1, we have introduced some main concerns with regard to privacy, the motivating projects of this thesis, as well as our objectives. In this chapter, we present the background on data anonymization techniques that our work relies on. We highlight that the content of this chapter is **primarily** inspired by existing literature in books [59, 110] and papers [20, 28, 108, 46]. Appropriate references to other works are provided throughout this chapter.

2.1/ Introduction: Syntactic VS Algorithmic Privacy

In the literature, many privacy models have been proposed to tackle privacy issues. In this manuscript, we consider two data privacy definitions, namely, *Syntactic privacy* and *Algorithmic privacy*. More specifically, the former notion tries to define a syntactic criterion that should be satisfied by the *output dataset* through transforming the data. The most influential method is named k-anonymity [18, 20], which was the starting point for other extensions like l-diversity [28] and t-closeness [29]. We introduce k-anonymity in Section 2.2, which will be used in Chapter 11. Throughout this manuscript, we will refer to **anonymity** as a condition of being "safe in the crowd" (i.e., anonymous).

The latter algorithmic notion considers that anonymization is a property of the *algorithm*, rather than the output dataset. This is the core insight of *differential privacy* [27, 26], which addresses the paradox of learning about a population while learning nothing about single individuals [59]. One special form of DP is the *non-interactive case* considered in this manuscript, which corresponds to, e.g., releasing summary statistics, the sanitized dataset, a synthetic dataset, and so on. Throughout this manuscript, we will refer to **sanitization** the fact that data anonymization was achieved through verifying DP (i.e., using a DP algorithm). In this manuscript, we consistently used differential privacy. So, we present the centralized model of DP in Section 2.3, the local model of DP in Section 2.4, and a local model of DP for location privacy in Section 2.5.

2.2/ k-ANONYMITY

Given a public medical database without identifiers but where age, ZIP code, ..., were present, and a $20\$$ dollars public voter records from Massachusetts, United States of America, a Ph.D. student named Latanya Sweeney was able to re-identify the Governor of Massachusetts in this medical database [72]. This re-identification attack took place because there was similar demographic information in both medical databases and voter list records. This way, the combination of several demographic data made people *unique* in both databases, which allowed Sweeney to directly match these records in both databases.

To tackle this *uniqueness* problem in data publishing, Samarati and Sweeney [18, 20] proposed the *k*-anonymity model, which requires that each released record to be indistinguishable from at least $k-1$ others. Intuitively, the larger *k* is the better the privacy protection will be. On applying *k*-anonymity, there is a difference between: *explicit identifiers* (e.g., names), which are removed or masked to avoid direct re-identification; *sensitive attributes* (e.g., disease), that might be preserved, and *quasi-identifiers (QIDs)* such as age and gender, in which *k*-anonymity seeks to ensure indistinguishability. We recall the definition of *k*-anonymity in the following.

Definition [1] (*k*-anonymity requirement [18, 20])**.** *Each release of data must ensure that every combination of values of QIDs can be indistinctly matched to at least k individuals.*

We also recall here an example from [28]. Table 2.1 exhibits a pseudonymized dataset (i.e., with no direct identifiers like 'name') that stores the medical record of a set of individuals. This dataset is composed of both sensitive (disease) and 'non-sensitive' information like age, gender, and nationality. Table 2.2 exhibits a 4-anonymous version of the original data in Table 2.1. Note that in Table 2.2, there is no *unique* record anymore and there are three different combinations of values grouped by $k = 4$ records.

However, several studies have pointed out limitations of the *k*-anonymity model, normally resulting in a new syntactic notion of privacy such as *l*-diversity [28] and *t*-closeness [29]. For instance, the last four records in Table 2.2 exhibits the same sensitive value *Cancer*. So, if an attacker with background knowledge knows someone within $[31;41[$ years old contributed to this dataset, it is obvious the disease value for this person. This is also known as *homogeneity* attack. Besides, *k*-anonymity does not *compose*, i.e., if the same person participates in two independent *k*-anonymous releases, there is no guarantee s/he will be *k*-anonymous in the composition of both dataset. Suppose the person in the first row (in red color) tested positive for tuberculosis in the hospital that release the 4-anonymous dataset of Table 2.2. Although this hospital had a good laboratory, the person decides to take a second test in another hospital, which releases the 5-anonymous dataset of Table 2.3. So, if an attacker knows, e.g., that someone is 29 years old, lives

	Quasi Identifiers – QIDs				Sensitive
ID	Zip	Age	Gender	Nationality	Disease
1	13053	28	M	Russian	Tuberculosis
2	13068	29	M	American	Heart
3	13068	21	F	Japanese	Viral
4	13053	23	M	American	Viral
5	14853	49	M	Indian	Cancer
6	14853	48	F	Russian	Heart
7	14850	47	M	American	Viral
8	14850	49	F	American	Viral
9	13053	31	M	American	Cancer
10	13053	37	M	Indian	Cancer
11	13068	36	F	Japanese	Cancer
12	13068	35	F	American	Cancer

Table 2.1: An example of a pseudonymized dataset (adapted from [28]).

Quasi Identifiers – QIDs				Sensitive	
Zip	Age	Gender	Nationality	Disease	
130**	[21;31[	*	*	Tuberculosis	4 individuals
130**	[21;31[	*	*	Heart	
130**	[21;31[	*	*	Viral	
130**	[21;31[	*	*	Viral	
148**	[41;50[	*	*	Cancer	4 individuals
148**	[41;50[	*	*	Heart	
148**	[41;50[	*	*	Viral	
148**	[41;50[	*	*	Viral	
130**	[31;41[	*	*	Cancer	4 individuals
130**	[31;41[	*	*	Cancer	
130**	[31;41[	*	*	Cancer	
130**	[31;41[	*	*	Cancer	

Table 2.2: A 4-anonymous dataset of Table 2.1 (adapted from [28]).

in ZIP code 13012, and visited both hospitals, the *unique* record that matches in both Tables 2.2 and 2.3 is the first one (also in red color). Thus, jeopardizing this user privacy since k-anonymity does not compose.

2.3/ DIFFERENTIAL PRIVACY

Consider a database that stores the result of an infectious disease of a set of individuals (e.g., Table 2.1). From this database, we could learn statistics about the underlying population and publish these statistics publicly. However, information might leak about specific individuals in the database, which could compromise their privacy. In theory, we would like that the global information relative to the population to be public, e.g., "how many people tested positive for this disease". At the same time, we would like that the

Quasi Identifiers – QIDs				Sensitive	
Zip	Age	Gender	Nationality	Disease	
130**	< 35	*	*	Tuberculosis	
130**	< 35	*	*	Diabetes	
130**	< 35	*	*	Parkinson	5 individuals
130**	< 35	*	*	Parkinson	
130**	< 35	*	*	Diabetes	
148***	≥ 35	*	*	Heart	
148***	≥ 35	*	*	Cancer	
148***	≥ 35	*	*	Viral	5 individuals
148***	≥ 35	*	*	Cancer	
148***	≥ 35	*	*	Cancer	

Table 2.3: An example of a 5-anonymous dataset from a second hospital.

information of each individual to be private, i.e., not releasing "*who* tested positive for the disease". Unfortunately, this is not always possible. For instance, if each time an attacker adds or removes someone of the database and performs the query "how many people tested positive for this disease?", in the end, it is possible to infer whose people tested positive by calculating the influence of each individual.

One way to preserve privacy in this scenario is to add some *noise* in the output of the query, which, *ideally*, should not destroy the utility of the data. In other words, the challenge would be to maximize the utility of the released noisy statistics while preserving the privacy of the individuals. Differential privacy (DP) [27, 26] is a formal definition that allows quantifying the privacy-utility trade-off. Indeed, rather than being a privacy property of the *output* dataset (like *k*-anonymity and its variants), DP is a definition that must be respected by a randomized *algorithm* (i.e., algorithmic notion of privacy).

In recent years, DP has been increasingly accepted as the current standard for data privacy with several large-scale implementations in the real-world [219] (cf. [132, 232, 95, 106, 196, 187, 61, 169, 220, 114, 205]). One key reason is that DP addresses the paradox of learning about a population while learning nothing about single individuals [59]. More specifically, the idea is that removing (or adding) a single row from the database should not affect *much* the statistical results. A formal definition of DP is given in the following.

Definition [2] ((ϵ, δ)-Differential Privacy [59])**.** *Given $\epsilon > 0$ and $0 \leq \delta < 1$, a randomized algorithm $\mathcal{A} : \mathcal{D} \to R$ is said to provide (ϵ, δ)-differential-privacy ((ϵ, δ)-DP) if, for all neighbouring datasets $D_1, D_2 \in \mathcal{D}$ that differ on the data of one user, and for all sets R of outputs:*

$$\Pr[\mathcal{A}(D_1) \in R] \leq e^{\epsilon} \Pr[\mathcal{A}(D_2) \in R] + \delta. \tag{2.1}$$

The additive δ on the right-side of Eq. (2.1) is interpreted as a probability of failure. Normally, a common choice for δ is to set it significantly smaller than $1/n$ where n is the

number of users in the database [59]. Throughout this manuscript, if $\delta = 0$, we will just say that $\mathcal{A}$ is ϵ-DP.

Notice that if ϵ (a.k.a. the *privacy loss* or the *privacy budget*) is zero, both distributions are equal, and in this case, there is no leakage of information. This is equivalent to the privacy goal stated by Dalenius [12] in 1977 as "*access to a statistical database should not enable one to learn anything about an individual that could not be learned without access*". However, respecting such a statement, as proven in [26], no utility could ever be obtained. So, we have to accept leaking *some* information about individuals in order to have some utility, which is translated to increasing ϵ (i.e., *privacy-utility trade-off*).

2.3.1/ PROPERTIES OF DIFFERENTIAL PRIVACY

Differential privacy possesses several important properties, highlighting its strength in comparison with other privacy models. For instance, with DP, there is no need to define the *background knowledge* that attackers might have, which is equivalent to assuming an attacker with *unlimited resources*. Besides, DP definition protects *anything* associated with a single individual, e.g., their presence in the database and their sensitive information [181]. On the other hand, DP does not protect against attribute inference as it may leak information about individuals **not** present in the database.

In addition, DP is immune to *post-processing*, which means it is not possible to make an ϵ-DP mechanism less differentially private by evaluating any function f of the response of the mechanism, given that there is no additional information about the database.

Proposition 1 (Post-Processing of DP [59]). *If $\mathcal{A} : \mathcal{D} \to R$ is ϵ-DP, then $f(\mathcal{A})$ is also ϵ-DP for any function f.*

Furthermore, DP also *composes* well, which is one of the most powerful features of this privacy model. For instance, accounting for the *overall* privacy loss consumed in a pipeline of several DP algorithms applied to the same database is feasible due to composition. We recall two types of composition below.

Proposition 2 (Sequential Composition [59]). *Let $\mathcal{A}_1$ be an ϵ_1-DP mechanism and $\mathcal{A}_2$ be an ϵ_2-DP mechanism. Then, the mechanism $\mathcal{A}_{1,2}(\mathcal{D}) = (\mathcal{A}_1(\mathcal{D}), \mathcal{A}_2(\mathcal{D}))$ is $(\epsilon_1 + \epsilon_2)$-DP.*

Proposition 3 (Parallel Composition [59]). *Let $\mathcal{A}_1$ be an ϵ_1-DP mechanism and $\mathcal{A}_2$ be an ϵ_2-DP mechanism. Let D_1 and D_2 be arbitrary disjoint subsets of the input domain $\mathcal{D}$. Then, the mechanism $\mathcal{A}_{1,2}(\mathcal{D}) = (\mathcal{A}_1(D_1), \mathcal{A}_2(D_2))$ is $max(\epsilon_1, \epsilon_2)$-DP.*

2.3.2/ Differentially Private Mechanisms: Laplace and Gaussian

Any mechanism that respects Definition 2 can be considered differentially private. Two widely used DP mechanisms for numeric queries (i.e., functions $f : \mathcal{D} \rightarrow \mathbb{R}$) are the Laplace mechanism [27] and the Gaussian mechanism [59]. One important parameter that determines how accurately we can answer the queries is their *sensitivity*. We recall the definition of ℓ_1- and ℓ_2-sensitivity and both Laplace and Gaussian mechanisms below, respectively.

Definition 3 (ℓ_1-sensitivity [59]). *The ℓ_1-sensitivity of a function $f : \mathcal{D} \rightarrow \mathbb{R}$, for all neighbouring datasets $D_1, D_2 \in \mathcal{D}$ that differ on the data of one user, is:*

$$\Delta_1(f) = max \, \|f(D_1) - f(D_2)\|_1$$

Definition 4 (ℓ_2-sensitivity [59]). *The ℓ_2-sensitivity of a function $f : \mathcal{D} \rightarrow \mathbb{R}$, for all neighbouring datasets $D_1, D_2 \in \mathcal{D}$ that differ on the data of one user, is:*

$$\Delta_2(f) = max \, \|f(D_1) - f(D_2)\|_2$$

Definition 5 (Laplace mechanism [27]). *For a query function $f : D \rightarrow \mathbb{R}$ over a dataset $D \in \mathcal{D}$, the Laplace mechanism is defined as:*

$$\mathcal{A}_L(D, f(.), \epsilon) = f(D) + Lap\left(\frac{\Delta_1}{\epsilon}\right),$$

in which $Lap(b)$ is the Laplace distribution centered around 0 and of scale b. The Laplace mechanism is proven to preserve ϵ-DP [27].

Definition 6 (Gaussian mechanism [59]). *For a query function $f : D \rightarrow \mathbb{R}$ over a dataset $D \in \mathcal{D}$ and for $\sigma = \frac{\Delta_2}{\epsilon}\sqrt{2\ln(1.25/\delta)}$, the Gaussian mechanism is defined as:*

$$\mathcal{A}_G(D, f(.), \epsilon, \delta) = f(D) + \mathcal{N}\left(0, \sigma^2\right)$$

in which $\mathcal{N}\left(0, \sigma^2\right)$ is the normal distribution centered at 0 with variance σ^2. For $\epsilon \in (0, 1)$, the Gaussian mechanism provides (ϵ, δ)-DP [59].

2.3.3/ Privacy amplification by sampling

There exist scenarios in which using a random subsample of the database is sufficient to approximate the overall distribution of the original database (e.g., census data). Sampling is a fundamental tool in the design of differentially private mechanisms as there is an

amplification effect [25, 43, 118, 173, 32]. For instance, amplification by sampling plays a key role in machine learning since many classes of algorithms utilize sampling strategies during the training process (e.g., differentially private stochastic gradient descent [73]).

So, why is there an amplification effect? Informally, assume we extract a random subsample from a database and, next, we apply a DP mechanism to this sampled database. Observe now that there is more *uncertainty* on the output of the DP mechanism since an attacker would be, first, unable to distinguish which data samples were used and, second, there is the DP guarantee. More rigourously, Li et al. [43, Theorem 1] theoretically prove this effect.

Theorem 1. ***Amplification by Sampling*** *[43]. Let $\mathcal{A}$ be an ϵ'-DP mechanism and $\mathcal{S}$ to be a sampling algorithm with sampling rate β. Then, if $\mathcal{S}$ is first applied to a dataset $\mathcal{D}$, which is later sanitized with $\mathcal{A}$, the derived result satisfies ϵ-DP with $\epsilon = \ln\left(1 + \beta(e^{\epsilon'} - 1)\right)$.*

2.4/ LOCAL DIFFERENTIAL PRIVACY

The centralized DP model from Section 2.3, assumes that a trusted curator has access to compute on the entire raw data of users. 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 [228]. To preserve privacy at the user-side, an alternative approach, namely, local differential privacy (LDP), was initially formalized in [32]. With LDP, rather than trusting in a data curator to have the raw data and sanitize it to output queries, each user applies a DP mechanism to their data before transmitting it to the data collector server. A formal definition of LDP is given in the following:

Definition 7 (ϵ-Local Differential Privacy)**.** *A randomized algorithm $\mathcal{A}$ satisfies ϵ-local-differential-privacy (ϵ-LDP) if, for any pair of input values $v_1, v_2 \in Domain(\mathcal{A})$ and any possible output y of $\mathcal{A}$:*

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

Intuitively, ϵ-LDP guarantees that an attacker can not distinguish whether the true value is v_1 or v_2 (input) with high confidence (controlled by ϵ) irrespective of the background knowledge one has. This is because both values have approximately the same probability to generate the same perturbed output. Similar to the centralized model of DP, LDP also enjoys the properties described in Section 2.3.1, e.g., immunity to post-processing and composition [59].

The LDP model allows collecting data in unprecedented ways and, therefore, has led to several adoptions by industry. For instance, big tech companies like Google, Apple, and

Microsoft, reported the implementation of LDP mechanisms to gather statistics in well-known systems (i.e., Google Chrome browser [61], Apple iOS and macOS [106], and Windows 10 operation system [95]). Indeed, there is a rich literature on LDP models [50, 123, 67, 217, 159, 243, 88, 61, 95, 108, 81, 80, 166, 83, 106, 139, 179, 150], and we refer the interest reader to recent survey works on LDP [206, 204, 209].

In this manuscript, **we focus on the *fundamental* problem of private frequency (or histogram) estimation** under ϵ-LDP guarantees. This is a primary objective of LDP, in which the data collector decodes all the sanitized data of the users and can then estimate the number of users for each possible value. The frequency estimation task has received considerable attention in the literature [217, 159, 108, 80, 139, 116, 61, 95, 168, 192, 150] as it is a building block for other complex tasks (e.g., heavy hitter estimation [67, 238, 88, 143], estimating marginals [161, 138, 134, 78], frequent itemset mining [136, 85]).

Let $A_j = \{v_1, v_2, ..., v_{c_j}\}$ be a set of $c_j = |A_j|$ values of a given attribute and let ϵ be the privacy budget. Each user u_i, for $i \in \{1, 2, ..., n\}$, has a value $v \in A_j$. Thus, the aggregator's goal is to estimate a c_j-bins histogram, including the frequency of all values in A_j. Algorithm 1 exhibits the general procedure for frequency estimation under LDP, which includes: `Encoding` and `Randomization` at the user-side, and `Aggregation` at the server-side (i.e., the aggregator).

Algorithm 1 General procedure for frequency estimation under LDP

Input : Original data of users, privacy parameter ϵ, and local randomizer $\mathcal{A}$.
Output : Estimated frequencies.

`# User-side`

1: **for** each user u_i ($i \in \{1, 2, ..., n\}$) with input value $v \in A_j$ **do**
2: `Encode`(v) into a specific format (**if needed**);
3: `Randomize`(v) with $\mathcal{A}(v, \epsilon)$;
4: Transmit the randomized output to the aggregator.
5: **end for**

`# Server-side`

6: The server `aggregates` the reported values and estimates their frequency.
7: **return :** c_j-bins histogram, including the frequency of all values in A_j.

In addition, if one intends to collect data from the same population, i.e., longitudinal studies, the authors in [61] introduced the concept of *memoization*. The idea behind *memoization* is to use two steps of sanitization, where the first step uses an upper bound value of ϵ_∞-LDP and only outputs lower epsilon reports using this randomized data. This will be a subject of study in Chapter 6. In the next three subsections, we will review state-of-the-art LDP protocols for non-longitudinal frequency estimation (a.k.a. frequency oracles).

2.4.1/ RANDOMIZED RESPONSE

Randomized response (RR) is a surveying technique proposed by Warner [11], to provide plausible deniability for individuals responding to embarrassing questions. Suppose we want to do a survey to know "how many people have already cheated on their partner". Due to social embarrassment, people would probably hesitate to answer this question honestly, thus lying on their answer. Instead, with RR, users would benefit from plausible deniability to their answers, following the scheme below.

Each user, throw a secret coin:

- If Tails throw the coin again (ignoring the outcome) and answer the question honestly;
- If Heads, then throw the coin again and answer "Yes" if Head, and "No" if Tail.

Notice that even if users might have answered "Yes", we still would not be sure if they answered honestly or at random. With more details, Figure 2.1 illustrates the probability tree of the RR protocol with an unbiased coin (i.e., with equal probability $1/2$).

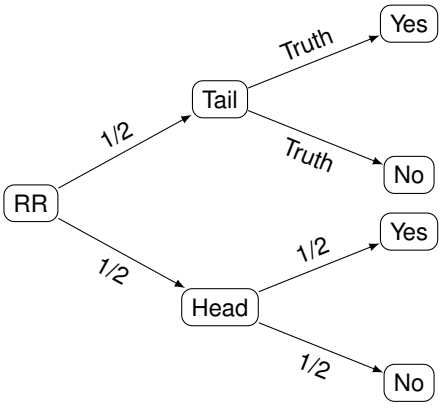

Figure 2.1: Summary of randomized response method with unbiased coins (i.e., with equal $1/2$ probability).

From Figure 2.1, let $\mathcal{A}$ represent the RR mechanism, we can calculate the following probabilities:

$$\Pr[\mathcal{A}(Yes) = Yes] = \Pr[\mathcal{A}(No) = No] = 0.75, \tag{2.2}$$

$$\Pr[\mathcal{A}(Yes) = No] = \Pr[\mathcal{A}(No) = Yes] = 0.25. \tag{2.3}$$

So, now, the objective is to estimate the frequency of "Yes" and "No" answers, i.e., the distribution of the original data. Let $f(v_y)$ be the proportion of *true* "Yes" answers and N_y

be the proportion of *observed* "Yes" answers. The following equation gives an estimated relation between these two variables:

$$N_y \approx \frac{1}{2} f(v_y) + \frac{1}{4} n.$$

The higher the number of samples n, with high probability, the more accurate the frequency estimation will be. In this case, $f(v_y)$ can be estimated with:

$$\hat{f}(v_y) \approx 2N_y - \frac{1}{2} n.$$

Similarly, we can calculate the number of estimated "No" answers. Translating the unbiased-coin RR model to DP theory, this model satisfies ϵ-LDP with $\epsilon = \ln\left(\frac{0.75}{0.25}\right) = \ln(3)$ [59]. More generically, given $v \in \{0, 1\}$ we can design an RR protocol to satisfy an arbitrary ϵ value (i.e., with biased coins) with the following perturbation function [80, 81]:

$$\forall y \in \{0, 1\} \Pr[\mathcal{A}_{RR(\epsilon)}(v) = y] = \begin{cases} p = \frac{e^\epsilon}{e^\epsilon + 1}, & \text{if } y = v \\ q = \frac{1}{e^\epsilon + 1}, & \text{if } y \neq v, \end{cases}$$

This satisfies ϵ-LDP since $\frac{p}{q} = e^\epsilon$. Notice that the RR algorithm does not require any encoding technique. To estimate the normalized frequency $f(v_i)$ that a value $v_i \in V$ occurs where $V = \{v_1, v_2\} = \{0, 1\}$, one calculates [80, 81]:

$$\hat{f}(v_i) = \frac{N_i - nq}{n(p - q)}, \tag{2.4}$$

in which N_i is the number of times the value v_i has been reported and n is the total number of users. In Theorems 1 and 2 from [108], it is shown that $\hat{f}(v_i)$ is an **unbiased estimation** of the true frequency $f(v_i)$ (i.e., $E[\hat{f}(v_i)] = f(v_i)$), and **the variance of this estimation** is calculated as:

$$Var[\hat{f}(v_i)] = \frac{q(1 - q)}{n(p - q)^2} + \frac{f(v_i)(1 - p - q)}{n(p - q)}. \tag{2.5}$$

Since the estimation in Eq. (2.4) is unbiased, its variance $Var[\hat{f}(v_i)]$ is equal to the mean squared error (MSE) [1] that is commonly used as an accuracy metric (e.g., cf. [202, 203, 239, 224]), also adopted throughout this manuscript. More formally,

$$\begin{aligned} MSE &= \frac{1}{|V|}\sum_{v\in V} E\left[\left(\hat{f}(v_i) - f(v_i)\right)^2\right] \\ &= \frac{1}{|V|}\sum_{v\in V}\left(Var[\hat{f}(v_i)] + (E[\hat{f}(v_i)] - f(v_i))^2\right) \\ &= \frac{1}{|V|}\sum_{v\in V} Var[\hat{f}(v_i)]. \end{aligned} \tag{2.6}$$

Furthermore, with no knowledge about the real frequency $f(v_i)$ and because in real life the vast majority of values appear very infrequently, we will consider $f(v_i) = 0$. Notice that this is common practice in the literature (e.g., cf. [108, 239]), which provides an approximation for the variance as [108]:

$$Var^*[\hat{f}(v_i)] = \frac{q(1-q)}{n(p-q)^2}. \tag{2.7}$$

Replacing $p = \frac{e^\epsilon}{e^\epsilon+1}$ and $q = \frac{1}{e^\epsilon+1}$ into Eq. (2.7), the RR variance is calculated as:

$$Var^*[\hat{f}_{RR}(v_i)] = \frac{e^\epsilon}{n(e^\epsilon-1)^2}.$$

2.4.2/ GENERALIZED RANDOMIZED RESPONSE

The *k*-Ary RR [80] mechanism extends RR to the case of $c_j \geq 2$ and it is also referred to as direct encoding [108] (since no particular encoding needed) or generalized RR (GRR) [136, 203, 138]. Throughout this manuscript, we will use the term GRR for this LDP protocol. Given a value $v \in A_j$, *GRR(v)* outputs the true value with probability p, and any other value $v' \in A_j$ such that $v' \neq v$ with probability $1-p$. More formally, the perturbation function is defined as:

$$\forall y \in A_j \Pr[\mathcal{A}_{GRR(\epsilon)}(v) = y] = \begin{cases} p = \frac{e^\epsilon}{e^\epsilon+c_j-1}, & \text{if } y = v \\ q = \frac{1}{e^\epsilon+c_j-1}, & \text{if } y \neq v. \end{cases}$$

GRR satisfies ϵ-LDP since $\frac{p}{q} = e^\epsilon$. The estimated frequency $\hat{f}(v_i)$ that a value v_i occurs for $i \in [1, c_j]$ is also calculated using Eq. (2.4). Replacing $p = \frac{e^\epsilon}{e^\epsilon+c_j-1}$ and $q = \frac{1}{e^\epsilon+c_j-1}$ into Eq. (2.7), the GRR variance is calculated as:

$$Var^*[\hat{f}_{GRR}(v_i)] = \frac{e^\epsilon + c_j - 2}{n(e^\epsilon-1)^2}. \tag{2.8}$$

2.4.3/ UNARY ENCODING PROTOCOLS

Protocols based on unary encoding (UE) consist of transforming a value v into a binary representation of it. So, first, for a given value v, $B = Encode(v)$, where $B = [0, 0, ..., 1, 0, ...0]$, a c_j-bit array where only the v-th position is set to one. Next, the bits from B are flipped independently, depending on parameters p and q, to generate a sanitized vector B', in which:

$$\Pr[B'_i = 1] = \begin{cases} p, \text{ if } B_i = 1 \\ q, \text{ if } B_i = 0. \end{cases}$$

The proof that UE-based protocols satisfy ϵ-LDP for

$$\epsilon = ln\left(\frac{p(1-q)}{(1-p)q}\right), \tag{2.9}$$

is known in the literature and can be found in [61, 108]. In [108] the authors presents two ways for selecting probabilities p and q, which determines the protocol variance. One well-known UE-based protocol is the Basic One-time RAPPOR [61], referred to as symmetric UE (SUE), which selects $p = \frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}$ and $q = \frac{1}{e^{\epsilon/2}+1}$, where $p + q = 1$ (symmetric). The estimated frequency $\hat{f}(v_i)$ that a value v_i occurs for $i \in [1, c_j]$ is also calculated using Eq. (2.4). Replacing $p = \frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}$ and $q = \frac{1}{e^{\epsilon/2}+1}$ into Eq. (2.7), the SUE variance is calculated as [61]:

$$Var^*[\hat{f}_{SUE}(v_i)] = \frac{e^{\epsilon/2}}{n(e^{\epsilon/2}-1)^2}. \tag{2.10}$$

Moreover, rather than selecting p and q to be symmetric, Wang et al. [108] proposed optimized UE (OUE), which selects parameters $p = \frac{1}{2}$ and $q = \frac{1}{e^{\epsilon}+1}$ that minimize the variance of UE-based protocols while still satisfying ϵ-LDP. Similarly, the estimation method used in Eq. (2.4) equally applies to OUE. Replacing $p = \frac{1}{2}$ and $q = \frac{1}{e^{\epsilon}+1}$ into Eq. (2.7), the OUE variance is calculated as [108]:

$$Var^*[\hat{f}_{OUE}(v_i)] = \frac{4e^{\epsilon}}{n(e^{\epsilon}-1)^2}. \tag{2.11}$$

2.4.4/ ADAPTIVE LDP PROTOCOL

Comparing Eq. (2.8) with Eq. (2.11), elements $c_j - 2 + e^{\epsilon}$ is replaced by $4e^{\epsilon}$. Thus, as highlighted in [108], when $c_j < 3e^{\epsilon} + 2$, the utility loss with GRR is lower than the one of OUE. This adaptive selection of LDP protocol has been used in many settings in the

literature [138, 136]. Throughout this manuscript, we will use the term adaptive (ADP) to denote this best-effort and dynamic selection of LDP mechanism.

2.5/ GEO-INDISTINGUISHABILITY

Geo-indistinguishability (GI) [46] is based on a generalization of DP developed in [48] and has been proposed for preserving location privacy without the need of a trusted curator (e.g., a malicious location-based service), i.e., a local DP model. A mechanism satisfies ϵ-GI if for any two locations x_1 and x_2 within a radius r, the output y of them is (ϵ, r)-geo-indistinguishable if we have:

$$\frac{\Pr(y|x_1)}{\Pr(y|x_2)} \leq e^{\epsilon r}, \forall r > 0, \forall y, \forall x_1, x_2 : d(x_1, x_2) \leq r.$$

Intuitively, this means that for any point x_2 within a radius r from x_1, GI forces the corresponding distributions to be at most $l = \epsilon r$ distant. In other words, the level of distinguishability l increases with r, e.g., an attacker can distinguish that the user is in Paris rather than London but can hardly (controlled by ϵ) determine the user's exact location. Although both GI and DP use the notation of ϵ to refer to the privacy budget, they cannot be compared directly because ϵ in GI contains the unit of measurement (e.g., meters).

On the continuous plane (as we consider in this manuscript), an intuitive polar Laplace mechanism has been proposed in [46] to achieve GI, which is briefly described in the following. Rather than reporting the user's true location $x \in \mathbb{R}^2$, we report a point $y \in \mathbb{R}^2$ generated randomly according to $D_\epsilon(y) = \frac{\epsilon^2}{2\pi} e^{-\epsilon d_2(x,y)}$. Algorithm 2 shows the pseudocode of the polar Laplace mechanism in the continuous plane. More specifically, the noise is drawn by first transforming the true location x to polar coordinates. Then, the angle θ is drawn randomly between $[0, 2\pi)$ (line 3), and the distance r is drawn from $C_\epsilon^{-1}(p)$ (line 5), which is calculated using the negative branch W_{-1} of the Lambert W function [5]. Finally, the generated distance and angle are added to the original location.

Algorithm 2 Polar Laplace mechanism in continuous plane [46]

Input : $\epsilon > 0$, real location $x \in \mathbb{R}^2$.
Output : sanitized location $y \in \mathbb{R}^2$.
1: Draw θ uniformly in $[0, 2\pi)$
2: Draw p uniformly in $[0, 1)$
3: Set $r = C_\epsilon^{-1}(p) = -\frac{1}{\epsilon}\left(W_{-1}\left(\frac{p-1}{e}\right) + 1\right)$
4: **return :** $y = x + \langle r\cos(\theta), r\sin(\theta)\rangle$

2.6/ CONCLUSION

In this chapter, we have revised state-of-the-art anonymization techniques. We started with the well-known k-anonymity model, presenting its definition, an intuitive example, as well as some of its limitations. We then presented differential privacy, which is a definition that should be satisfied by a randomized algorithm. While the former satisfies a syntactic notion of privacy, i.e., the final database should satisfy "k-anonymity", DP is a property of the process. In addition, DP offers strong post-processing and composition properties, which are important in designing differentially private systems for real-life applications. Besides, we have presented the decentralized setting of DP, also known as local DP, in which there is no need to assume a trusted server. In the LDP setting, the aggregator already knows the users' identifiers, but not their private data. In this case, users apply a differentially private algorithm in their own device such that only perturbed data is sent to the aggregator. Also, we have presented geo-indistinguishability, which is an LDP model to protect location privacy. Geo-indistinguishability utilizes a Laplacian noise to perturb the actual location of a user before transmitting to the (un)trusted server and has received considerable attention due to its effectiveness and simplicity of implementation (e.g., Location Guard [6]). Lastly, for each DP model, we have presented the main mechanisms that will be used throughout this manuscript.

3

Machine Learning and Databases Used on Experiments

In Chapter 2, we have revised the background on data anonymization techniques. In this chapter, we now briefly review the background on machine learning techniques and concepts that our work utilizes, as well as the databases we experiment on. We highlight that the content of this chapter related to machine learning is **primarily** inspired by existing literature [151, 79, 51]. Appropriate references to other works are provided throughout this chapter.

3.1/ Introduction to Machine Learning

Following the definition of machine learning (ML) given by Géron in their book [151] "*Machine Learning is the science (and art) of programming computers so they can learn from data*." In contrast with traditional programming techniques that are based on conditional and loop statements, ML automatically learns *from data*. The way of *learning* ranges, e.g., from supervised, unsupervised, semi-supervised, and reinforcement learning. In this manuscript, we focus only on **supervised** learning, in which the ML algorithms also receive the desired outputs (e.g., a scalar, a label). ML supervised applications typically solve prediction and classification tasks both approached in this manuscript.

3.1.1/ Classification Problems

Classification predictive modeling problems have as main goal to predict a class label. Indeed, based on a set of input X the objective is to classify each sample in a given discrete label y. The output variables are frequently referred to as labels or categories. A classical example of a classification task is spam filters, in which a classifier is trained over emails labeled as spam or not spam. In general, depending on the objective one may

want to train ML classification algorithms for binary, multiclass, or multilabel problems, for example.

3.1.2/ Regression Problems

Regression predictive modeling problems have as main goal the prediction of a numerical value. More precisely, based on a set of input X, the objective is to predict a numerical value y. For example, the price of a house may be predicted by using as predictors the number of bedrooms, its area, its location, and so on. Generally, a problem with multiple inputs is often referred to as a multivariate regression problem. One special type of regression is with **ordered data**, also known as *time-series* data. In these cases, the order of the samples *matters*. Indeed, time-series data is a set of observations collected by repeated measures throughout time. There are many practical applications for time series data in both classification and regression problems. For example, forecasting the spread of infectious diseases [231], tracking financial market indices [199], and forecasting human mobility [225], to name a few.

3.1.3/ Modeling Techniques

To select the most performing ML algorithm per problem we tackled, we generally evaluated one or more among the ML models described in the next three subsections (Section 3.1.3.1– 3.1.3.3).

3.1.3.1/ Linear Model

In this manuscript, we only considered a regularized version of the Linear Regression model, namely, least absolute shrinkage and selection operator (LASSO) [15], which is widely used for prediction purposes. The LASSO is a method of contracting the coefficients of the regression, whose ability to select a subset of variables is due to the nature of the constraint on the coefficients. Originally proposed by Tibshirani [15] for models using the standard least squares estimator, it has been extended to many statistical models such as generalized linear models. We used the LASSO implementation from the Scikit-learn library [36].

3.1.3.2/ Decision Tree Algorithms

One of the popular predictive modeling techniques used in ML is decision tree learning [90]. Decision tree-based algorithms are often chosen for predictive modeling be-

cause of their interpretability and high performance. We evaluated two decision tree learning algorithms in this manuscript:

- Extreme Gradient Boosting (XGBoost) [76] is a decision-tree-based ensemble ML algorithm that produces a predictive model based on an ensemble of weak predictive models (decision trees). XGBoost uses a novel regularization approach over standard gradient boosting machines, which significantly decreases the model's complexity. The system is optimized by a quick parallel tree construction and adapted to be fault-tolerant under distributed environments.

- Light Gradient Boosted Machine (LGBM) [100] is a novel gradient boosting framework, which implemented a leaf-wise strategy. This strategy significantly reduces computational speed and resource consumption in comparison to other decision tree-based algorithms.

3.1.3.3/ Artificial Neural Networks

Another popular active research area in ML is artificial neural networks. Neural networks are the foundation of deep learning (DL), which has become a progressively popular research topic. We used the Keras library [68] to implement all our DL models. Throughout this manuscript, we will evaluate one or more of the following DL methods:

- Multilayer Perceptron (MLP) is an artificial neural network of the feedforward type [79, 69, 92], characterized by a unidirectional flow of computation. MLPs are based on the interconnection of several units (neurons) to transmit signals, which are normally structured into three or more layers, namely, input, hidden(s), and output.

- Recurrent neural network (RNN) is a specialized class of neural networks used to process sequential data (e.g., time-series data). RNNs have at least one feedback connection that provides the ability to use contextual information when mapping between input and output sequences. In this manuscript, we have applied three state-of-the-art improvements over the standard RNN, which are described in the following:

  - Long Short-Term Memory [16] is a type of RNN that overcomes the vanishing gradient problem of standard RNNs. Inside its cell memory unit, the learning process is controlled by three gates: input, forget, and output, which give LSTM the ability to learn which data in a sequence is important to keep or to discard.

- Gated Recurrent Unit [57] is also a type of RNN, which works using the same principle as LSTM. GRU utilizes two gates: update and reset, which decide what information should be passed to the output.
- Bidirectional RNN (BiRNN) [17] is a combination of two RNNs: one RNN moves forward while the other moves backward. That is, BiRNN connects two hidden layers of opposite directions to the same output. The RNN cells in a BiRNN can either be standard RNNs, LSTMs, GRUs, and so on.

3.1.4/ Model Selection and Hyperparameter Tuning

Generally, besides multiple alternatives of ML algorithms for a given task, there are as well several hyperparameters to tune in each of them. More precisely, let be given the definition from [51]: "*The process of evaluating a model's performance is known as model assessment, whereas the process of selecting the proper level of flexibility for a model is known as model selection.*"

Throughout this manuscript, we assess the performance scores of our ML models on a *hold-out* testing set. In the following two subsections, we describe the performance metrics (Section 3.1.5) and the hyperparameters' optimization methods (Section 3.1.6) considered in this manuscript.

3.1.5/ Performance Metrics

Throughout this manuscript, we used common metrics from the literature to evaluate the performance of ML models. For **regression** tasks, we considered using one or more of the following metrics:

- Root mean squared error (RMSE) measures the square root average of the squares of the errors and is calculated as: $RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$;
- Mean absolute error (MAE) measures the averaged absolute difference between real and predicted values and is calculated as: $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$;
- Mean absolute percentage error (MAPE) measures how far the model's predictions are off from their corresponding outputs on average and is calculated as: $MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%$;
- Coefficient of determination (R^2) measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s);

in which y_i is the real output, $\hat{y}_i$ is the predicted output, and n is the total number of samples, for $i \in [1, n]$. In addition, for **binary classification** tasks, we considered the following metrics:

- Accuracy (ACC) measures how many observations, both positive and negative, were correctly classified.
- *Recall* measures how many observations out of all positive observations have been classified as positive.
- *Precision* measures how many observations predicted as positive are indeed positive.
- Macro average F1-Score (MF1) is the harmonic mean between precision and recall with macro average, which calculates metrics for each label and finds their unweighted mean.

3.1.6/ HYPERPERAMETER OPTIMIZATION

The goal of hyperparameter optimization in ML is to discover the set of hyperparameters of a particular ML algorithm that returns the best performance measured on a *hold-out* set. The search space defines the volume to be searched, with each dimension being a hyperparameter and each point representing a model configuration. In this manuscript, we mainly used Bayesian optimization (BO) [86] and random search optimization [40]. On the one hand, to apply a random search, one initially defines the search space as a bounded domain of hyperparameters values. Next, each step of the optimization randomly samples a point in that domain, builds the model, and then evaluates its performance. In the end, the random search optimization selects the most accurate method encountered during the iterative process. On the other hand, in contrast to random search, Bayesian methods track the entire set of prior evaluations of hyperparameters, which are used to build a probabilistic model of mapping hyperparameters to the likelihood of a score of an objective function. Rather than random sampling points in the domain, the goal of BO is to improve as iterations go by.

3.2/ MACHINE LEARNING WITH DIFFERENTIAL PRIVACY

In this manuscript, we consider two differentially private ML settings, which depend on where the DP guarantee is added. As revised in Section 2.3.1, DP is immune to post-processing, which means that after the differentially private step, everything stays DP [59]. The two considered settings are described in the following two subsections.

3.2.1/ Differentially Private Input Perturbation

Input perturbation (or data perturbation) consists to the fact that **DP is added to each data sample $\mathbf{x}_i \in \mathcal{D}$**. For example, let $\mathbf{x}$ be a real-valued vector, then a differentially private version of it using the Laplace mechanism (cf. Section 2.3.2) is: $\hat{\mathbf{x}} = \mathbf{x} + Lap(b)$. This is also true for categorical data, e.g., by randomizing each data point in $\mathbf{x}$ with some LDP protocol (i.e., frequency oracle) from Section 2.4. On the one hand, input perturbation is the easiest method to apply [54, 140] and it is independent of any ML and post-processing techniques. On the other hand, the perturbation of each sample in the dataset may have a negative impact on the utility of the trained model.

In the literature, some works [191, 98] started to investigate how 'input perturbation' through applying the Gaussian mechanism [59] on data samples can guarantee (ϵ, δ)-DP on the final ML model. In [211], the authors sanitized each sample with LDP protocols (GRR [80] for categorical data and the Piecewise mechanism [166] for real-valued data) for training ML models to compare with federated learning. Indeed, there are an extensive literature on training ML models over differentially private data (e.g., [124, 170, 93, 243, 179, 60, 210, 172, 148]).

3.2.2/ Differentially Private Gradient Perturbation

Another solution to guarantee DP to the trained model is perturbing intermediate values in iterative algorithms. In Chapter 8 of this manuscript, we considered training deep learning models with DP guarantees. In this case, the authors in [73] proposed a differentially private version of the stochastic gradient descent algorithm (DP-SGD). Indeed, DL models trained with DP-SGD provide provable DP guarantees for their input data. Two new parameters are added to the standard stochastic gradient descent algorithm, namely, *clip* and *noise multiplier*. The former is used to bound how much each training point can impact the model's parameters, and the latter is used to add controlled Gaussian noise to the clipped gradients in order to ensure DP guarantee to each data sample in the training dataset. There are many works in differentially private DL literature (e.g., [131, 241, 155, 71, 145, 230, 102]).

3.3/ Presentation of Databases Used on Experiments

This section presents the databases shared by the OBS team (Section 3.3.1), the pre-processed SDIS 25 datasets resulting of the work carried out by our collaborator Selene Cerna (Section 3.3.2), and open datasets from the UCI ML [96] repository (Section 3.3.6).

3.3.1/ FLUX VISION MOBILITY REPORTS

The first motivating project of this manuscript concerns multidimensional CDRs-based mobility reports released by OBS throughout time. On the one hand, the OBS team initially shared a database of **daily statistics for a single area** (Section 3.3.1.1). We used this first database in Chapter 4 with the main goal of improving the utility of these data. In addition, the OBS team provided us with a more informative database of 30**-minutes statistics for six areas of interest** (Section 3.3.1.2). We used this second database in Chapter 8 with the main goal of evaluating the privacy-utility trade-off of differentially private DL models on a multivariate time series forecasting task.

3.3.1.1/ TOURISM MOBILITY REPORTS

One important use case of CDRs has been to analyze the mobility patterns of people in tourist events [194, 62, 53]. The first database at our disposal, from now on named **FIMU-DB**, regards multiple *tourism statistics* on the frequency of visitors *by days and by the union of consecutive days*. OBS considered 'visitors' people present at least 1 hour between 06:00 and 23:59 of a given day of the reporting period in the area of interest. The geographical space is the area of an international music festival named "Festival International de Musique Universitaire" (FIMU). The FIMU is organized and financed by the City of Belfort, France, with the support of student associations. The 31^{st} edition of the FIMU occurred on the first five days of June 2017 [94].

The FIMU-DB has seven different files. Among them, *five files* describe for each day, the cumulative number of unique visitors on the last Nb days, where Nb ranges from 1 to 7 days. These files are labeled from now on as FO_country, FR_geo, FR_Gender, FR_region, and FR_age, where 'FO' stands for foreigners and 'FR' stands for French citizens.

In each file relating to French citizens, people are grouped according to their visitor category. "Resident" are people whose billing address is the administrative area around the FIMU. "French tourist" are people billed in France but not in the aforementioned category. The FO_country file has only people grouped as "Foreign tourist" who are people with a foreign mobile phone operator.

In summary, each file aggregates people according to the **cumulative count** from 1 to 7 days (i.e., *the number of people in the union of consecutive days*), and also by specific categories, which are briefly detailed below:

1. The FR_Gender file contains 3,776 rows at total and distinguishes the people by **gender** (masculine, feminine, and Not Registered – NR). Furthermore, during the analysis, we noticed very few differences in the frequency of men and women per

day (about 50% for both). Hence, in this study, the NR values were changed to masculine or feminine, with an equal probability of 50%;

2. The FR_age file contains 8,820 rows at total and groups the visitors by **age ranges** as: '<18', '18-24', '25-34', '35-44', '45-54', '55-64', '>65', and 'NR';

3. The FR_geo file contains 14,989 rows and groups the visitors in a specific category named **geolife**, divided into different socio-professional sub-categories as: 'NR', 'comfortable family pavilion', 'traditional rural', 'comfortable family urban', 'secondary residence', 'popular', 'dynamic rural', 'growing peri-urban', 'rural worker', 'dynamic urban', 'middle-class urban', and 'low-income urban';

4. The FR_region file contains 50,350 rows and groups the visitors in the specific category named (French) **region** as: 'AUTRE 97', 'Centre', 'Languedoc-Roussillon', "Provence-Alpes-Côte d'Azur", 'Lorraine', 'Ile-de-France', 'Franche-Comté', 'Midi-Pyrénées', 'Corse', 'Basse-Normandie', 'Aquitaine', 'Poitou-Charentes', 'Pays de la Loire', 'Nord-Pas-de-Calais', 'Champagne-Ardenne', 'Bourgogne', 'Bretagne', 'Alsace', 'Rhône-Alpes', 'Picardie', 'Auvergne', and 'Haute-Normandie';

5. The FO_country file contains 10,832 rows and groups the foreign visitors by **country** as: 'Belgium + Luxembourg', 'Asia Oceania', 'Netherlands', 'Scandinavia', 'United Kingdom', 'Italy', 'Spain', 'China', 'Other countries in Europe', 'Germany', 'United States', 'Russia', 'Swiss', 'Eastern country', and 'Rest of the world'.

For instance, Table 3.1 exhibits 5 random samples to illustrate how the volume data are grouped by geolife profiles in the FR_geo file. In addition, Fig. 3.1 illustrates the cumulative number of people for the three first consecutive FIMU's days using the same FR_geo file (randomly replacing # values for an integer within 1 and 20).

Table 3.1: Number of unique visitors per geolife present on days of FIMU.

Date	Geolife	Visitor category	Cumulative days	Volume
2017-06-01	comfortable family pavilion	French Tourist	7 days	2751
2017-06-02	low-income urban	Resident	4 days	3355
2017-06-03	comfortable family pavilion	Resident	3 days	# (i.e., <20)
2017-06-04	secondary residence	French Tourist	1 days	97
2017-06-05	rural worker	Resident	3 days	1,359

Furthermore, **the remaining two files** labeled from now on as Nights_actual and Presence_time. Unlike previous data files, these latter files **do not** consider cumulative days information, but the volume of visitors each day $(Nb = 1)$. Similarly, both files classify the data by the main categories (Resident, French tourist, Foreign tourist) and by specific categories described below:

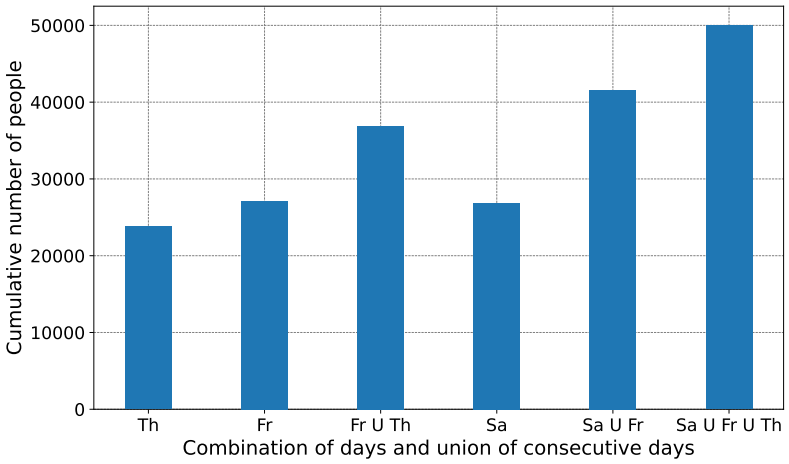

Figure 3.1: Cumulative number of people for the three first consecutive days of FIMU, i.e., for Thursday (Tu), Friday (Fr), and Saturday (Sa). For instance, Sa U Fr means the **union** of Saturday and Friday.

- The Nights_actual file has 1,145 rows describing for each day the number of visitors who spent a night at the relevant date. Here, people are grouped by a specific category namely **sleeping area** where people spent the night. There sleeping areas are: 'Agglomeration of Hericourt', 'Rest Territory of Belfort', 'NR', 'City of Belfort', 'Vosges', 'Rest of Doubs', 'Rest of Haute Saone', 'North Haut Rhin', 'Agglomeration of Belfort', 'Agglomeration of Montbeliard', and 'South Haut Rhin';

- The Presence_time file has 1,301 rows describing for each day the number of hours where visitors were present in the area of interest. Here, people are grouped by a specific category namely **visit duration** within several sub-categories as: 'Duration 2h', 'Duration 3h', 'Duration 4h', 'Duration 5h', 'Duration 6h', 'Duration 7h', 'Duration 8h', 'Duration 9h', 'Duration 10h', and 'Duration 10h-18h'. For instance, 'Duration 2h' matches people present between one and two hours.

We noticed that in the Nights_actual file, the total volume of visitors per day is much less compared to the previous five files (around 4,000 on average). This means that many people did not spend the night near the city of Belfort. Therefore, considering the number of visitors per day from all other files and those in Nights_actual, the term NR was assigned to people that did not sleep in the area of interest.

3.3.1.2/ GEOMARKETING REPORTS

Another use case of CDRs is understanding people mobility during the spread of infectious diseases [156, 65, 197, 186, 174]. The second database at our disposal regards multiple published Flux Vision [53] statistics for geomarketing purposes, which **were col-**

lected during the novel Coronavirus Disease 2019 (COVID-19) pandemic [247, 201] in 2020. The complete database *is fully available online* in [180].

We only used the file named "**presence30min.csv**", which comprises information for two periods: from 2020-04-20 to 2020-05-03 and from 2020-08-24 to 2020-11-04. This dataset has frequency statistics by 30 minutes (min) on the number of users by "Zone" (i.e., 6 regions in Paris) and by "type" (i.e., French or foreign). The geographical space (i.e., Zone) concerns 6 specific regions in Paris, France, named "Commune Montreuil", "IRIS 930480204", "IRIS 930480205", "IRIS 930480206", "IRIS 930480401", and "IRIS 930480604" in the original file.

We applied the following preprocessing to the original "presence30min.csv" file. We aggregated the number of users by "type" for each of the 6 regions, i.e., *focusing only on the total number of users per the* 6 *regions*. In addition, for each week, region, and 30-min interval, we used the interquartile range technique [4] to detect outliers and missing data. These values were completed with the average value for that respective week, region, and 30-min interval. We will refer to this pre-processed dataset as **Paris-DB** throughout this manuscript.

More formally, the Paris-DB is a multivariate time series dataset $X_{(t_1,t_\tau)}$ with aggregate number of people per 6 regions and corresponding time period $t \in [1, \tau]$ of 30-min intervals. That is, $X_{(t_1,t_\tau)} = [\langle t_1, \mathbf{x}_1 \rangle, \langle t_2, \mathbf{x}_2 \rangle, ..., \langle t_\tau, \mathbf{x}_\tau \rangle]$, where $\mathbf{x}_t$ is a vector of size 6 in which each position represents the number of users per region at time $t \in [1, \tau]$.

On analyzing the Paris-DB, Fig. 3.2 illustrates the **total number of people** for two 14-days periods: from the beginning of 2020-04-21 to the end of 2020-05-03 and from the beginning of 2020-09-23 to the end of 2020-10-06. The plot on the left-side corresponds to mobility analytics during the first national lockdown period in France [2] because of the COVID-19 pandemic. The plot on the right-side corresponds to a period with no lockdown measures. As one can notice, there is a clear difference between the first period of analysis (low mobility activity) and the second one (high mobility activity). This type of mobility analysis provides important insights on mobility patterns for public authorities and policymakers to fight the COVID-19 pandemic, for example [237, 218].

3.3.2/ FIREMEN DATABASE

As mentioned in Chapter 1, the AND team has been investigating ML-based solutions to optimize the SDIS 25 (i.e., and EMS in France) services. Our connection with the SDIS 25 is through the Ph.D. student Selene Cerna, with a CIFRE thesis (N 2019/0372) and a strict confidentiality agreement to use SDIS 25 original data. In this section, we present three datasets processed by her. These datasets have also been used by Selene Cerna to develop the ML-based solutions with original data that we use for comparison purposes,

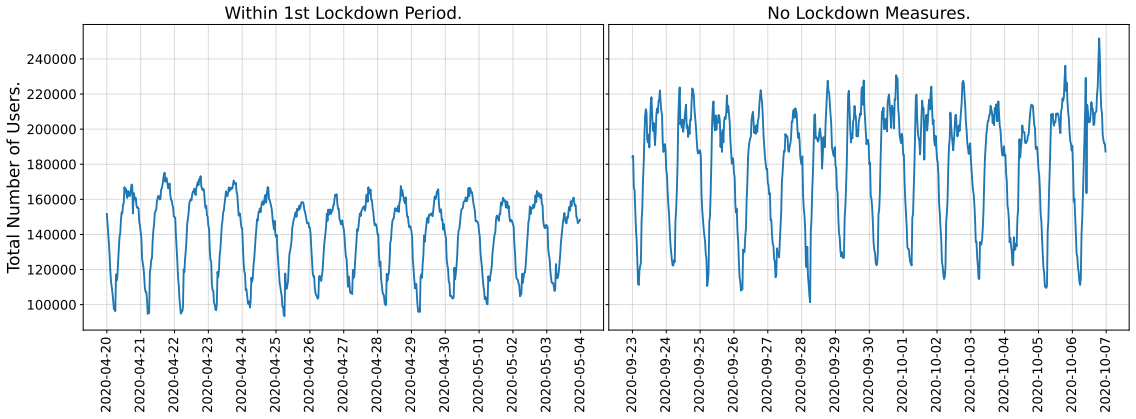

Figure 3.2: Aggregated human mobility analytics during two weeks within the first lockdown period in France (left-side plot) and during two weeks with no lockdown measures (right-side plot).

i.e., to evaluate the **privacy-utility trade-off** of our solutions.

3.3.3/ Interventions Data

Predicting the operational demand of EMS is a way to allow their reallocation of human and material resources (e.g., cf. [75, 152, 146, 177, 176, 226]). So, the first dataset we use, from now on named **Interv-DB**, has information about 382046 **interventions** attended by the SDIS 25 from 2006 to 2018 in 608 cities inside the Doubs region. The **Interv-DB** has two attributes: *SDate*, which is the precise "*Starting Date*" of the intervention, and the *City* where the intervention took place. The way this dataset has been preprocessed will be explained in Chapter 9, as part of a proposed privacy-preserving methodology that allows both statistical learning and forecasting tasks of firemen demand by region.

3.3.4/ Response Time Data

Although predicting the operational demand may help on the redeployment of resources, another solution would be to predict the response time of each ambulance, which would allow, e.g., to move from a static resource deployment plan into a dynamic one. The second dataset we use contains information about 186130 **dispatched ambulances** from SDIS 25 centers that attended 182700 EMS interventions from 2006 up to June 2020. After a preprocessing step carried out by Selene Cerna, the final dataset, from now on named **ART-DB**, has the following attributes:

- **Temporal features.** Based on the time the SDIS 25 has been notified, a few temporal features were included, such as the: year, month, day, weekday, hour, and

categorical indicators to denote holidays, end/start of the month, and end/start of the year;

- **Operation demand features.** The number of interventions attended by the SDIS 25 in the *past hour* and the number of *active* interventions in the current hour;
- **Traffic data.** These are prediction of traffic level for the Doubs region as indicators ranging from 1 (regular flow) to 4 (extremely difficult flow) per day from [244];
- **Weather data.** These are historical weather information from [246] such as precipitation, temperature, wind speed, gust speed, and so on, which were added according to the hour of each intervention;
- **Location-based features.** The latitude and longitude coordinates of the intervention and of the SDIS 25 center that took charge of the intervention; the district, the city, and the zone of the intervention;
- **Computed features.** The great-circle distance [3] between the SDIS 25 center and the emergency scene; the estimated travel time, and the estimated driving distance. These two latter features were obtained with the open source routing machine (OSRM) [35] API;
- **The scalar target variable** is the ambulance response time (ART) in minutes, which is the time measured from the SDIS 25 notification to the ambulance's arrival on the emergency scene.

3.3.5/ Calls, Victims, and Operators Data

From another point of view, identifying high urgent situations (i.e., a life-or-death situation) would allow EMS to quickly respond to victims needing priority attention (i.e., if they might die). Therefore, our third dataset, from now on named **Vic_Mort-DB**, has information about 177883 **victims** that the SDIS 25 attended from January 2015 to December 2020. After a preprocessing step carried out by Selene Cerna, the **Vic_Mort-DB** has the following attributes:

- **Victim data.** The victim's age, gender, and city (where the intervention occurred);
- **(Call center) operator data.** The operator's age, gender, grade (e.g., commander, captain), and seniority (i.e., experience time in days);
- **Temporal features.** Based on the time the SDIS 25 has been notified, a few temporal features were included, such as the: hour, day, day of the week, month, and year;

- **Call/Intervention data.** The delay time to answer the phone, total call duration, delay time to diffuse the alert (i.e., to notify an SDIS 25 center), the SDIS 25 center that assisted the victim, and the type of intervention. The latter is described by 3 variables: type of operation (e.g., aid to person, fire), the subtype of operation (e.g., an emergency, fire on the public road, fire in an individual room), and the motive for departure (e.g., cardiac arrest, respiratory distress);
- **Calculated features.** Probability of mortality by *motive* and by *age*, which are calculated according to the learning set only; the age of the victims grouped into 8 categories; the total sum of delay time to answer the phone, call duration, and delay time to diffuse the alert; and the great-circle distance between the center and the city;
- **The target variable** is the victim's mortality, which is a binary attribute (0: alive, 1: dead).

3.3.6/ Open Datasets

For ease of reproducibility of the works carried out in Chapters 6 and 7, we (also)[1] considered **three multidimensional open datasets from the UCI ML repository** [96]. These datasets were selected because they allow evaluating our solutions more practically, i.e., with typically real-world datasets. For instance, they differ on the number of users n, on the number of attributes d, on the number of values per attribute $\mathbf{c} = [c_1, c_2, ..., c_d]$, and on the data distribution of each attribute. These datasets are described in the following.

- *Nursery*[2]*.* This dataset contains $n = 12960$ samples and $d = 8$ categorical attributes. The domain size of each attribute is $\mathbf{c} = [3, 5, 4, 4, 3, 2, 3, 3, 5]$, respectively.
- *Adult*[3]*.* This dataset contains 48842 samples extracted from the 1994 Census database. There are 14 attributes (including the income attribute), in which 9 are categorical and 5 are numerical (i.e., considering 'education' instead of 'education-num'). After removing all samples with missing values (i.e., symbol '?'), there are $n = 45222$ samples in this dataset. We only selected the categorical attributes (i.e, $d = 9$). The domain size of each attribute is $\mathbf{c} = [7, 16, 7, 14, 6, 5, 2, 41, 2]$, respectively.
- *Census-Income*[4]*.* This dataset contains weighted census data from the 1994 and 1995 years. There are 40 demographic and employment related variables (including the total person income attribute), in which 33 are categorical and 7 are numerical. In total, there are $n = 299285$ samples in this dataset. We only selected

[1] Besides the multidimensional open dataset generate in Chapter 4.
[2] https://archive.ics.uci.edu/ml/datasets/nursery
[3] https://archive.ics.uci.edu/ml/datasets/adult
[4] http://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29

the categorical attributes (i.e, $d = 33$). The domain size of each attribute is $\mathbf{c} = [9, 52, 47, 17, 3, 7, 24, 15, 5, 10, 2, 3, 6, 8, 6, 6, 51, 38, 8, 10, 9, 10, 3, 4, 5, 43, 43, 43, 5, 3, 3, 3, 2]$, respectively.

3.4/ CONCLUSION

In this section, we have revised state-of-the-art ML techniques and some concepts. We started revising supervised learning and classification and regression tasks, all three considered in this manuscript. Next, we have reviewed state-of-the-art ML algorithms ranging from linear (i.e., LASSO), decision-tree learning (i.e., LGBM and XGBoost), and deep learning (e.g., MLP, RNNs) models. We also presented the metrics that will be used to assess the models' performance, as well as two hyperparameter tuning methods (i.e., random search and Bayesian optimization). Lastly, we also reviewed how to build differentially private ML models, which fundamentally depends on where the DP guarantee is added. That is, by the post-processing property of DP [59] (cf. Section 2.3.1), everything after DP, stays DP. Indeed, we mainly consider in this manuscript the rigorous input perturbation setting, which sanitizes each data sample independently (i.e., row-by-row). Although utility may drop, we believe the privacy-utility trade-off is worthwhile since the sanitized dataset will be protected if data leakes [228], and the ML model will also be differentially private, protecting the data against, e.g., data reconstruction attacks, membership inference attacks [105, 104]. On the other hand, we also briefly presented another setting, namely, gradient perturbation contextualized to deep learning methods trained with the DP-SGD [73] algorithm. Lastly, we also presented the datasets we will be using in this manuscript to perform our experiments.

III

Contribution: Improving the Utility and Privacy of LDP protocols

4

MS-FIMU: A Multidimensional Dataset to Evaluate LDP protocols

In Chapters 1, 2, and 3, we have presented all main components that will be used in the rest of the contribution chapters of this manuscript. In this chapter, we start to study statistics on aggregated human mobility data generated by OBS (i.e., with the Flux Vision system [53]). The main objective here is to instantiate a mobility scenario from these statistics and to generate a synthetic dataset that allows simulating the data collection pipeline with the privacy-preserving techniques we develop in the next three Chapters 5, 6, and 7. We emphasize that although the **exact** Flux Vision's anonymization method is **unknown** to the author, we refer to state-of-the-art methods that could give similar results.

4.1/ Introduction

The main objective of this chapter is to propose an approach to instantiate a mobility scenario that matches the **anonymized** dataset of mobility described in Section 3.3.1.1 named FIMU-DB, which was published by OBS. To generate the FIMU-DB, as stated by OBS, algorithms for data acquisition are compliant with European laws to guarantee the **anonymity** of each person. Indeed, following the GDPR [112] and CNIL [13], MNOs must anonymize "on-the-fly" CDRs used for purposes other than billing. More precisely, if CDRs are used for mobility analytics, these data must be processed within a required time interval (e.g., 15 minutes) if and only if there is a sufficient number of users present for reaching a specific level of anonymity (i.e., "hide in the crowd"). Besides, all implicit metadata (e.g., users' IDs, timestamps) should be excluded before transmitting any data for processing. That is to say, data should be aggregated (respecting anonymity) within the required time interval and all kinds of identifiers must be excluded before any further

analysis.

More specifically, to generate FIMU-DB, OBS established pre-defined indicators through generalization (e.g., age ranges, socio-professional profiles, ...). Next, an anonymity threshold of $k = 20$ was defined, i.e., if there are less than 20 users the number is masked with the symbol #. Besides, given the number of identified Orange customers, an extrapolation algorithm was developed to estimate the real population. This latter algorithm can be seen as a perturbation-based technique to add noise to the true value. Thus, within the required time interval, OBS processed CDRs respecting $k = 20$ for any interval to produce the mobility indicators per day and per the **union** of consecutive days. We notice that to generate cumulative statistics, i.e., the number of unique users by the union of days could have been done, e.g., using Bloom filters [22]. Also, we use k to indicate the anonymity threshold as it follows the "hide in the crowd" protection provided by *k*-anonymity [18, 20]. However, in our view, we believe that the OBS procedure approximates some of the privacy-preserving approaches described in [89].

In summary, the FIMU-DB is subject to noise resulting from the extrapolation of detected Orange clients and from the anonymization procedure to respect the GDPR and the CNIL. Besides, the FIMU-DB has information on the number of people present by the **union** of consecutive days (also referred to as 'cumulative' information throughout this chapter).

Therefore, on the one hand, with our proposed approach, we aim to improve the utility of this data, providing the number of people present by all the **intersections** of days. The mobility scenario we propose represents an invaluable source of information to the city public administration and private companies. Rather than being limited to the number of unique people present in certain regions per union of consecutive days, the scenario allows knowing if they are the same visitors or different visitors over the analyzed time period. With such specific information, companies and public administration would be able to manage their employees and equipment resources efficiently to improve accommodation and transportation systems according to peoples' mobility, thus providing better attendees comfort and security.

Besides, we propose to recreate the instantiated mobility scenario with virtual humans, such that the synthetic dataset matches the original statistical data. Therefore, as an open dataset, one can carry out studies such as testing and improving data sanitization techniques. **For the rest of this manuscript, we will refer to the final synthetic dataset as Mobility Scenario FIMU (MS-FIMU)**, which contains 7 categorical attributes for $88,935$ unique users along 7 days (on average $\sim 26,000$ unique users per day). That is to say, a **longitudinal** and **multidimensional** dataset. **We invite the interested reader to visit the Github page (https://github.com/hharcolezi/OpenMSFIMU) to access the final results of this chapter and the published synthetic open dataset.**

The rest of this chapter is organized as follows.Section 4.2 presents the study case and

some challenges we faced working with real-world anonymized data. Section 4.3 introduces the proposed approach to instantiate a precise mobility scenario and to generate synthetic data. Section 4.4 presents the results and its discussion. Finally, Section 4.5 provides concluding remarks. The methodology presented in Section 4.3 and the results in Section 4.4 were published in a full paper [171] at the 16th International Wireless Communications & Mobile Computing Conference (IWCMC 2020).

4.2/ Study Case and Data Analysis

The main background for this chapter is the database named FIMU-DB from Section 3.3.1.1, which has seven main files: FR_gender, FR_age, FR_geo, FR_region, FO_country, Nights_actual, and Presence_time. In this section, we present the scenario in which OBS collected the data and we highlight some challenges one can face working with real-life anonymized data.

4.2.1/ Study Case

As reviewed in Section 3.3.1.1, the FIMU-DB was published by OBS, which collected statistics on the frequency of users on days and union of consecutive days through analyzing mobile phone data (i.e., CDRs). The geographical space is the area of an international music festival a.k.a "Festival International de Musique Universitaire" (FIMU).

Modeling people's mobility in such events is of great importance for public administration and private companies. Hence, we propose to model a more precise mobility scenario, including **one day before the FIMU event, the five days of the FIMU, and one day after the FIMU end. In other words, this $Nb = 7$ days scenario for a 5-days event provides information for these institutions to know the number of people who got in and out of the zone of analysis before, during, and after the event.**

4.2.2/ Challenges with Anonymized Statistical Data

Although the data produced by OBS are adequate for marketing purposes, conducting scientific studies using these data leads to, in our view, two challenges. First, we are unable to determine the real number of people when OBS published #, i.e., due to the anonymity threshold $k = 20$. On the one hand, one could think of excluding all # values, which would probably compromise the utility of the data. So, in this chapter, instead of excluding this information, we considered the option of randomly replacing # by an integer within the known range from 1 to 20. However, both solutions (excluding or randomly

assigning an integer) result in different cardinalities between the seven files that describe the same population, which represents an **inconsistency**.

For instance, Table 4.1 summarizes the records of the first three days of the FIMU. In this scenario, the first day of analysis is Thursday and it has only one record labeled as 'Th1'. Friday has two records labeled as 'Fr1' (only Friday) and 'Fr2' (Friday **OR** Thursday), respectively. And Saturday has three records labeled as 'Sa1' (only Saturday), 'Sa2' (Saturday **OR** Friday), and 'Sa3' (Saturday **OR** Friday **OR** Thursday), respectively. **For the rest of this chapter, we will be using this notation (e.g., 'Fr1', 'Fr2', ...) to indicate the 'cumulative' information (i.e., the union of consecutive days).**

In Table 4.1, both 'FR_geo', 'FR_region', and 'FR_age' columns present the total number of *unique* French visitors aggregated in each file. This is according to the 'Cum. days' attribute exemplified in Table 3.1 and after replacing randomly all # values. The same procedure is reproduced for the other files. **Theoretically, the information from all three columns 'FR_geo', 'FR_region', and 'FR_age' should be equal as they describe the same population. However, this is not true, and the difference between files changes depending on the replacement of all # values (unknown).**

Table 4.1: Unique French visitors present over three FIMU's days.

Label	Cum. days	FR_geo	...	FR_region	FR_age
Th1	01 day	23,816	...	23,598	23,810
Fr1	01 day	27,145	...	26,945	27,143
Fr2	02 days	36,917	...	36,758	36,915
Sa1	01 day	26,894	...	26,699	26,868
Sa2	02 days	41,615	...	41,373	41,589
Sa3	03 days	50,024	...	49,823	49,999

4.3/ PROPOSED APPROACH

Our goal is to improve the understanding of people's mobility behavior from the number of unique visitors per day and cumulative days to find out the number of unique visitors by the intersection of days. The whole proposed approach is summarized with a flowchart depicted by Fig. 4.1. In this particular study, the ultimate goal is to infer the number of people who stayed in the city for one or any combination of days (i.e., the aggregate number of users in each intersection of days), considering one week, including the FIMU event. Further, once the whole mobility scenario is known, the objective is to generate samples to build a synthetic dataset with virtual people. The approach is detailed and applied in the following two subsections.

4.3.1/ MOBILITY SCENARIO MODELING

As described on the left side of Fig. 4.1, first, we input data with cumulative information and replace the # values. A Boolean map is used to describe every combination of $Nb = 7$ consecutive days resulting in $2^{Nb} = 128$ variables. Then, each of the $Nb(Nb + 1)/2 = 28$ cumulative days is described as a Boolean vector with 0 (excluded) and 1 (included) values per combination of days according to the representative map.

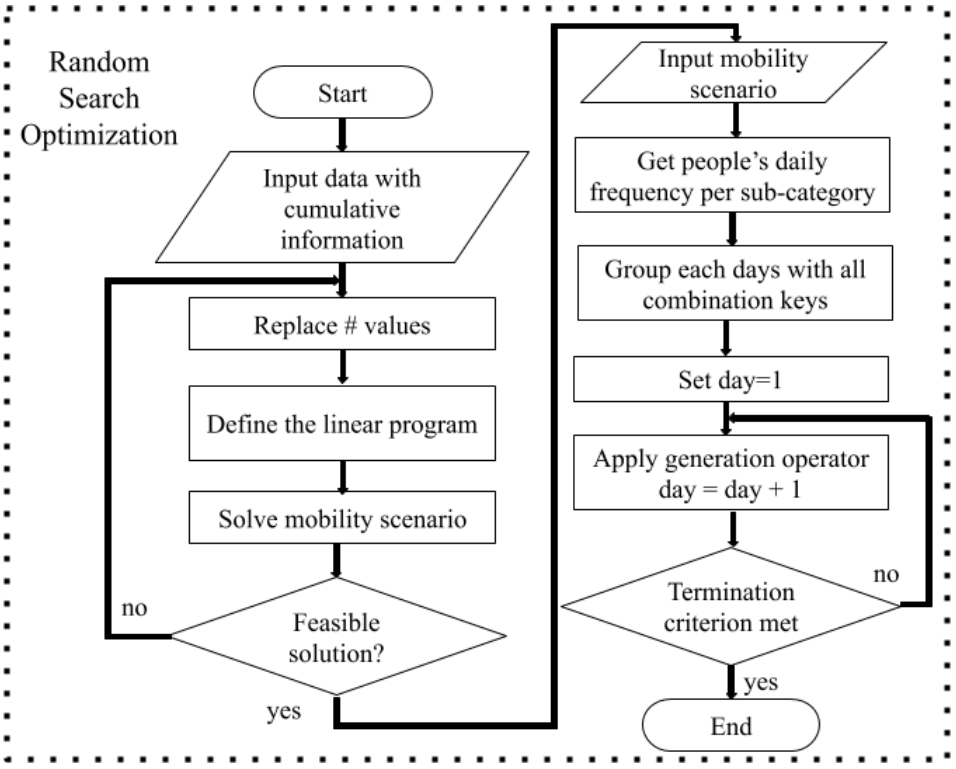

Figure 4.1: Flowchart of the proposed algorithm to, first, instantiate a mobility scenario with information by the intersection of days and, second, to generate synthetic data.

Then, a linear program (LP) is defined to instantiate the first feasible solution given a random initial solution, without trying to maximize or minimize any combination of days. The system constraints are the number of people per cumulative day, expressed as Boolean vectors. However, due to both problems of # values and **inconsistencies between the cardinalities of the datasets**, rather than using the exact 'known values', these problems are addressed by establishing bounds. In this case, the lower and upper bounds are the minimum and maximum values between all the datasets. The motivation for such an approach is to instantiate a feasible solution that respects the values of all available data such that the global error could minimize. More precisely, we are defining a linear constraint solver that computes an arbitrary solution within the set of feasible solutions rather than using the linear program as an optimization mechanism. In this case, the objective function of this system is just a constant (zero). Eq.(4.1) mathematically describes

the LP as:

$$\begin{aligned} \min \quad & 0, \\ \text{s.t.} \quad & lb_i \leq A_{ij} x_j \leq ub_i \\ & x_j \geq 0 \end{aligned} \tag{4.1}$$

$\forall i \in [1, Nb(Nb+1)/2]$ and $\forall j \in [1, 2^{Nb}]$ where A_{ij} is the Boolean matrix representing the Boolean vectors i and its respective days combinations j; x_j is the number of people per combination of days; and lb and ub are both lower and upper bounds, respectively, which corresponds to the total number of unique users.

Hence, instantiating a feasible solution for all the categories (Resident, French tourists, and foreign tourists) and grouping them as a unique mobility scenario provides the number of people for each combination of days. Then, with such results, the second part is retaken for generating samples of virtual humans aiming to approximate the original data.

Before moving on to step 2 (generate synthetic data), let us consider the scenario of Table 4.1 to better understand the proposed LP. Fig. 4.2 illustrates the Boolean map representation of $Nb = 3$ consecutive days (Th=Thursday, Fr=Friday, Sa=Saturday, and its complements), and the example of both $Th1$ (unique visitors on Thursday) and $Sa2$ variables (unique visitors present on Saturday **OR** Friday). Notice that $x1$ represents the number of visitors that are present neither Thursday nor Friday nor Saturday. This number is obviously not known and, hence, it is not considered.

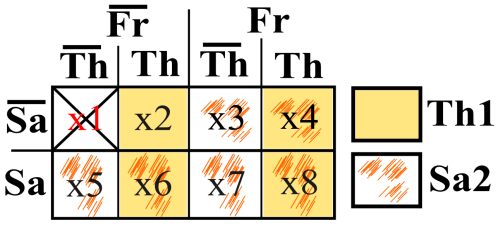

Figure 4.2: Representation of Nb=3 days combination and illustration of both Th1 and Sa2 known values.

Considering the LP in Eq.(4.1), Eq.(4.2) exhibits the A_{ij} matrix according to Fig. 4.2 and its lower (lb) and upper bound (ub) with values from Table 4.1 relating to French citizens.

$$\begin{bmatrix} Th1 \\ Fr1 \\ Fr2 \\ Sa1 \\ Sa2 \\ Sa3 \end{bmatrix} \Longrightarrow \begin{bmatrix} 23,598 \\ 26,945 \\ 36,758 \\ 26,699 \\ 41,373 \\ 49,823 \end{bmatrix} \leq \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_7 \\ x_8 \end{bmatrix} \leq \begin{bmatrix} 23,816 \\ 27,145 \\ 36,917 \\ 26,894 \\ 41,615 \\ 50,024 \end{bmatrix} \tag{4.2}$$

4.3.2/ SYNTHETIC DATA GENERATION

The proposed algorithm illustrated on the right side of Fig. 4.1 is summarized in the subsequent steps. First, using the original data, the frequency of visitors present each day of the week under study is calculated for each sub-category, e.g., on the first day 50.2% are men and 49.8% are women.

Next, we set up a dictionary for each day grouping its related keys of combination days; people present only Thursday are described by TT, people present both Thursday and Friday are described by TF, and so on. It is noteworthy that the same TF key appears on both Thursday and Friday dictionaries as they are the same people that attended both days in the analysis area.

Then, an iteration starts filling up each key for the first day with virtual individuals respecting the frequency of men and women, geolife categories, age groups, regions (countries for foreign tourists), the sleeping area, and the visit duration. Afterward, for the next six days, people with similar keys are directly copied from one day to another. In this case, the frequency for each category is re-calculated considering the existing people. The remaining people are then generated according to the new frequency. However, there is one exception about the attribute 'visit duration', which means that people could be present more hours from one day to another (a dynamic attribute). Hence, the approach may vary the duration time of every people each day relative to the real frequency acquired from the original data.

Once the stop criterion is met, i.e., when all days have their respective virtual humans, the error is calculated by querying the generated data and comparing it to the original one. The error, total error, and error rate metrics are defined in the following.

Definition [8] (Error). *Let $|A|$ be the cardinality of set A and $A_{|j}$ be the subset of A restricted to sub-category j, i.e., $A_{|j} = \{x | x \in A, x \in j\}$. Given a set O (original data), a set G (generated data), and sub-categories j related to each specific category (i.e., from the gender*

category there are two sub-categories, feminine and masculine), the error is defined as

$$error(j) = ||G_{|j}| - |O_{|j}||.$$

Definition 9 (Total Error)**.** *The total error TE is the sum of errors per sub-category j and per day i defined as*

$$TE = \sum_{i=1}^{n} \sum_{j=category} error(j)_i.$$

Definition 10 (Error Rate)**.** *The error rate ER is calculated considering j original datasets*

$$ER = \frac{TE}{\sum_{j=dataset} \sum_{j=category} |O_{j|j}|}.$$

These computations are repeated for m iterations based on a **random search optimization** approach. In particular, the first parameter randomly generated is the # values within the range 1-20, which changes the number of people per day and, consequently, per file at every iteration. In addition, considering the LP in Eq.(4.1), an initial solution is randomly generated such that the linear constraint solver can provide a different mobility scenario at each iteration. Then, the error rate metric is calculated. Finally, the mobility scenario and synthetic dataset with the smallest error rate are recovered as a final solution.

Some motivations for such a random search approach are described as follows. First, an initial attempt to model our problem as a linear program resulted in an infinity number of solutions. And second, as aforementioned, the # problem due to privacy constraints had to be handled, resulting in different cardinalities for the datasets.

4.4/ Results and Discussion

To carry out this work, we used the Pyeda Python package [245] for Boolean algebra operations. We applied the mixed-integer nonlinear programming solver from the Gekko package [119] to the proposed mobility scenario in Eq.(4.1). The Faker package [42] assigned fake French names for French citizens and default names (United States) for foreign tourists. All algorithms were implemented in Python 3. In order to run our codes, we used a machine with Intel (R) Core (TM) i7-8650 CPU @ 1.90GHz and 32GB RAM using Debian 10. In the next two subsections, we present our results.

4.4.1/ Mobility scenario

The random search algorithm performs $5,000$ evaluations of $m = 100$ iterations in parallel to search for the most representative distribution of people over the week of interest.

This is a suitable way to ensure a convergence pattern towards a global minimum due to probabilistic properties. At the end of 22 minutes, the random search stops, and the dataset is recovered with an error rate less than 8.1% at evaluation $1,050$ and iteration 79.

We summarize the final mobility scenario in Table 4.2, which presented the smallest error rate. In Table 4.2, each day of the week is represented in an abbreviated format, e.g., Sunday – Su and its complement by $\overline{Su}$. Besides, Fig. 4.3 depicts the decreasing error rate function based on the number of iterations. Lastly, for illustration purposes, Table 4.3 presents the number of visitors for both real and synthetic datasets (FR_age) and the absolute error for three sub-categories of ages on the first day of interest.

Days combination			$\overline{Fr}$				Fr			
			$\overline{Th}$		Th		$\overline{Th}$		Th	
			$\overline{We}$	We	$\overline{We}$	We	$\overline{We}$	We	$\overline{We}$	We
$\overline{Tu}$	$\overline{Mo}$	$\overline{Su}$ $\overline{Sa}$	-	4851	4378	1527	1801	1701	786	3450
		$\overline{Su}$ Sa	4791	234	87	266	1748	48	417	893
		Su $\overline{Sa}$	9695	228	199	508	341	92	506	1220
		Su Sa	2171	287	74	73	4237	103	1109	1229
	Mo	$\overline{Su}$ $\overline{Sa}$	5937	183	49	207	97	36	67	233
		$\overline{Su}$ Sa	592	100	103	42	63	116	63	80
		Su $\overline{Sa}$	7380	71	34	56	71	89	77	22
		Su Sa	256	51	96	49	27	52	94	61
Tu	$\overline{Mo}$	$\overline{Su}$ $\overline{Sa}$	7052	446	213	787	1163	35	679	775
		$\overline{Su}$ Sa	441	59	104	71	62	94	106	99
		Su $\overline{Sa}$	1004	110	53	70	85	87	52	53
		Su Sa	42	94	50	91	93	38	51	36
	Mo	$\overline{Su}$ $\overline{Sa}$	159	309	72	325	442	67	396	94
		$\overline{Su}$ Sa	111	76	89	35	71	34	102	434
		Su $\overline{Sa}$	434	84	71	41	112	67	89	149
		Su Sa	4176	61	71	93	211	74	506	**176**

Table 4.2: Final result of using our methodology, which produces a mobility scenario with the frequency of users per day and per the intersection of days.

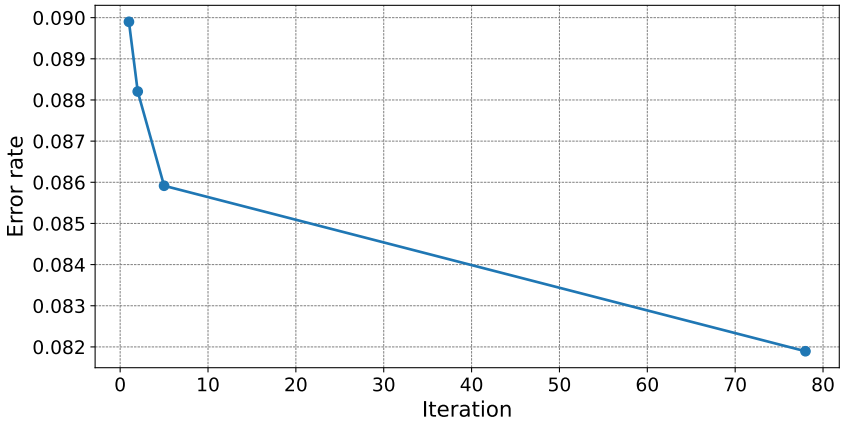

Figure 4.3: Decreasing error rate function through iterations.

Table 4.3: Number of visitors per dataset and absolute error for each sub-category of ages on the first day.

Age group	Real data	Synthetic data	Absolute Error
18-24	2,312	2,319	7 (0.3%)
35-44	3,230	3,215	15 (0.46%)
> 65	3,483	3,439	44 (1.26%)

4.4.2/ SYNTHETIC DATA

In the end, an open dataset is proposed with an associative table whose primary key is (Person ID, Date ID) combination, which specifies the visit duration information, as shown in Table 4.4. These two individual keys are linked to two other tables. The first, exemplified by Table 4.5, contains specific information about people, for instance, fake French names, geolife profile, and region. The second table maps the days under analysis, from the first to the last day respectively as follows: $\{1$: 2017-05-31$\}$, ..., $\{7$: 2017-06-06$\}$.

Table 4.4: Final format of the generated dataset.

Index	Person ID	Date ID	Visit Duration
1	5385	2	6h
2	234	5	4h

Table 4.5: Table with personal information about individuals.

Person ID	Name	Gender	Age	...	Visitor category	Region
91	Adrien Clement	M	45-54	...	French tourist	Alsace
32947	Grégoire Didier	M	25-34	...	French tourist	Franche-Comté
53990	Marie Le Lemaitre	F	25-34	...	Resident	Franche-Comté
58664	Michelle-Céline Marion	F	25-34	...	Resident	Franche-Comté

We recall here the information about all the attributes of the MS-FIMU dataset below.

- Static:
  - **Visitor Category** with 3 values: 'Resident', 'Foreign tourist', and 'French tourist';
  - **Gender** with 3 values: 'M' (masculine), 'F' (feminine), and 'NR' (Not Registered, e.g., for foreign people);
  - **Age** with 8 values: '<18', '18-24', '25-34', '35-44', '45-54', '55-64', '>65', and 'NR';
  - **Geolife** with 12 values: 'NR', 'comfortable family pavilion', 'traditional rural', 'comfortable family urban', 'secondary residence', 'popular', 'dynamic rural',

'growing peri-urban', 'rural worker', 'dynamic urban', 'middle-class urban', and 'low-income urban';

- **Region** with 37 values (countries for foreign people): 'Belgium + Luxembourg', 'Asia Oceania', 'Netherlands', 'Scandinavia', 'United Kingdom', 'Italy', 'Spain', 'China', 'Other countries in Europe', 'Germany', 'United States', 'Russia', 'Swiss', 'Eastern country', 'Rest of the world', 'AUTRE 97', 'Centre', 'Languedoc-Roussillon', "Provence-Alpes-Côte d'Azur", 'Lorraine', 'Ile-de-France', 'Franche-Comté', 'Midi-Pyrénées', 'Corse', 'Basse-Normandie', 'Aquitaine', 'Poitou-Charentes', 'Pays de la Loire', 'Nord-Pas-de-Calais', 'Champagne-Ardenne', 'Bourgogne', 'Bretagne', 'Alsace', 'Rhône-Alpes', 'Picardie', 'Auvergne', and 'Haute-Normandie';
- **Sleeping Area** with 11 values: 'Agglomeration of Hericourt', 'Rest Territory of Belfort', 'NR', 'City of Belfort', 'Vosges', 'Rest of Doubs', 'Rest of Haute Saone', 'North Haut Rhin', 'Agglomeration of Belfort', 'Agglomeration of Montbeliard', 'South Haut Rhin'.

- Dynamic:
  - **Visit Duration** with 10 values: 'Duration 2h', 'Duration 3h', 'Duration 4h', 'Duration 5h', 'Duration 6h', 'Duration 7h', 'Duration 8h', 'Duration 9h', 'Duration 10h', 'Duration 10h-18h'.

The motivation to release the synthetic open dataset with an associative table is to facilitate its improvement through adding more information about the population. Therefore, the associative table will remain unaltered, while more attributes can be added to the table with specific information about people. The generated dataset is available for anyone to freely access, use, modify, and share for any purpose at the aforementioned Github page (https://github.com/hharcolezi/OpenMSFIMU).

4.4.3/ Discussion and Related Work

In the literature, several studies on human mobility show that humans follow particular patterns with a high probability of predictability [49]. Hence, there is a high interest in understanding how people move. However, taking into account users' privacy, research emerges using synthetic and open data to solve such a problem. For example, in [99], the authors provided an approach for creating an open people mass movement dataset. In [74], the authors studied the use of open data for building and validating a realistic urban mobility model. The authors in [101] developed a framework for the generation of individual human mobility trajectories with realistic Spatio-temporal patterns. Finally,

the authors in [130] proposed a mobility dataset generation method of social vehicles traveling.

All the aforementioned works treated a different problem from ours, which corresponds to different data types available they have. In our case, there were only statistical mobility indicators with information about the unique number of people per day and per the union of consecutive days. We then proposed a solution based on linear programming (linear constraint solver) to instantiate a feasible solution and, thus, reconstruct virtual humans based on statistics. We also notice that the authors in [87] used a similar linear constraint solver to their problem.

Regarding our solution described in Section 4.3 and summarized in Fig. 4.1, one can notice that we have split the problem into two steps. Indeed, solving a single linear program considering the number of intersections $2^{Nb} = 128$ for each sub-category (e.g., masculine or feminine) of each category (e.g., gender) would probably require a large number of variables and, thus, it was not considered in this chapter. In addition, we consider that virtual humans have 'static' values for five attributes, i.e., each person has always the same geolife profile; people are from one unique region, they normally sleep in the same zone, and so on. The exception is for the 'Visit duration' attribute, which was considered 'dynamic' since people can vary the number of hours they stay in the FIMU per day.

Hence, as noticed in Fig. 4.3 and Table 4.3, the error metrics are very low when querying people in each sub-category from the generated data, compared to the original one. In other words, the result, which is one of many possible scenarios, closely describes how people behave during the week of interest. With such results, it is possible to assert with a reasonable amount of accuracy how many people were present in each combination (**intersection**) of 7 days, which is a more precise mobility scenario than just knowing the number of unique people per day or cumulative days (**union**).

For instance, from the final mobility scenario, and by querying the generated dataset, we can find out how many foreign tourists, French tourists, and residents are present only one or several days at the FIMU event, as well as their specific information such as socio-professional profile, region or countries, age groups, gender, and so on. For illustrative purposes, it is possible to know that from 176 visitors present during all week (see Table. 4.2, highlighted in bold), 153 are residents, 20 are French tourists, and 3 are foreign tourists.

Moreover, it is noticed that foreign tourists were present normally at one unique day or at most three consecutive days, which is consistent with reality. Indeed, foreign people come to the FIMU for few days and usually have no 'gaps' between days, such as one day present, the other not, and the next yes. Additionally, the premise of assigning one unique sleeping area for each visitor seems to indicate that the approach is consistent.

Such specific information on human mobility would be valuable for local communities and for accommodation and transportation companies, which would allow them to learn how people behave during a time period in a particular area. For instance, if one has information about the presence of foreign tourists on a specific combination of days and if they do not change much their sleeping place, accommodation companies can improve their future strategies to assist this population. Similarly, tourism companies would be more prepared knowing that most of the people present during the week are residents while tourists are rather present during the weekend of the FIMU event.

4.5/ CONCLUSION

This chapter proposes an approach to infer and recreate synthetic data that provides a precise mobility scenario based on one-week statistical data of **unions of consecutive days** made available by [53]. Our improved mobility scenario presents specific information about people present on one or several combinations of days (i.e., **all intersections of days**). The proposed approach is generic enough to apply to other mobility scenarios that rely on databases with information about the cumulative number of unique people for days (i.e., the union of consecutive days). Moreover, the proposed approach overcomes challenges due to data acquisition with anonymization techniques such as an anonymity threshold ($k = 20$ in this case) and extrapolation algorithms.

The results show that the proposal can be efficiently applied to generate a synthetic dataset with specific information about people present in a certain region, for instance, attending the FIMU [94] as was the case in this study. Finally, the generated and open dataset named MS-FIMU closely matches the original statistics with a low error rate, which substantiates the proposed approach. One direct **use case of MS-FIMU** is to evaluate differentially private cardinality estimation methods on longitudinal studies (e.g., [87, 66, 115]). Besides, **experimenting with machine learning tasks** could also be considered. Lastly, one can also **evaluate the effectiveness of new LDP protocols** on multidimensional and longitudinal frequency estimates, as we present in Chapters 6 and 7, or other complex tasks such as marginal estimation (e.g., [233, 161, 138, 134, 78]).

5

LDP-Based System to Generate Mobility Reports from CDRs

In Chapters 1, 3, and 4, we have reviewed mobility reports published by OBS Flux Vision system [53]. These mobility reports are, in other words, longitudinal statistics releases about the frequency of visitors by multiple attributes (e.g., as in [169, 62] too). Although current data privacy legislations require *anonymity* "on-the-fly" to collect CDRs for human mobility analytics, we posed ourselves two questions: **Q_1) what if** MNOs do not trust the data analyzers (e.g., third party companies working on mobility analytics)? Or **Q_2) what if** future data privacy legislations require different privacy protections than "anonymity on-the-fly", e.g., demanding "**sanitization** on-the-fly"? Indeed, while the former *anonymity* protection provides syntactic privacy through the "safe in the crowd" concept, the latter *sanitization* protection provides algorithmic privacy by using a DP model, as we defined in Section 2.1 for this manuscript. These questions $\mathbf{Q}_1$ and $\mathbf{Q}_2$ initially motivated the core contributions of this chapter where we propose an LDP-based CDRs processing system to generate mobility reports, following the objective of the OBS Flux Vision system. We invite the reader to refer to Chapter 2 for the background on LDP. Lastly, we highlight that although we refer to our proposal as *LDP-based*, this is a centralizer data owner (i.e., MNOs) that applies the LDP protocol on its servers, thus, providing ϵ-DP guarantees for users.

5.1/ Introduction

We start by recalling some requirements of data privacy regulations on collecting and analyzing CDRs for human mobility analytics. For instance, although MNOs have the right and duty to store CDRs, according to the GDPR [112], it does not mean MNOs have the right to use the collected *raw data* for other purposes. Besides, the CNIL [13], in France, requires that CDRs used for human mobility analysis (i.e., for purposes other than billing) to be *anonymized on-the-fly* (i.e., "hide in the crowd"). One reason behind

this is because users cannot sanitize their data locally since CDRs are automatically generated on MNOs' servers through the use of a service (e.g., making/receiving phone calls). Lastly, MNOs cannot process data for mobility analytics containing users' IDs or a hashed version of them as they are still unique IDs.

The purpose of this chapter is, thus, to propose a privacy-preserving system for human mobility analytics through mobile phone CDRs. This way, MNOs can benefit from such an important data source while respecting their clients' privacy and following major recommendations of data protection authorities such as the GDPR and CNIL. Indeed, we intend to analyze human mobility through *multidimensional* and *longitudinal* statistical data releases (e.g., as in [169, 62, 53]). Throughout this paper, "multidimensional" refers to data with multiple $d > 1$ attributes. For instance, as shown in Section 3.3.1.1 and Chapter 4, from subscription data, MNOs can gather: gender, age range (date of birth), and county origin (invoice address). From CDRs there is the coarse location (antennas that handled the service) and if it is "roaming data" (foreign mobile) or not. Besides, "longitudinal" refers to data with temporal information, i.e., analyzing human mobility over time (a.k.a. *continuous monitoring* in software literature [61, 95, 149]).

Therefore, we propose an LDP-based CDRs processing system for such purpose that goes beyond *"anonymity on-the-fly"*, i.e., with *"sanitization on-the-fly"*. The main reason to use the local DP model is that it allows sanitizing each sample independently while providing strong privacy guarantees (see Section 2.4). So, while MNOs CDRs processing systems utilizes each users' raw information (e.g., gender masculine, age range <18, ...), we propose that an LDP version of each users' data be used instead, i.e., LDP(masculine), LDP(<18). In fact, we assume that MNOs CDRs processing systems have, first, pre-defined the mobility indicators (e.g., gender, age-ranges, nationality, ...) they want to release.

Fig. 5.1 illustrates our system model overview and the considered trust boundary (in dashed line). The first entity, namely, **users**, refers to MNOs' clients, which are not able to sanitize their data locally when using a service (e.g., exchanging SMS). The second entity is the MNOs themselves, which are **data holders** and must ensure *"sanitization on-the-fly"* through an LDP mechanism *for all CDRs used for analyzing human mobility*. The third entity is the **data processor** (considered as an *untrusted* analyzer in our model), which processes data to generate statistical indicators. The last entity is the **data consumers**, which have access only to released statistics.

With more details, each time a user makes a call, or sends SMS, or connects to the internet ..., a CDR is generated and is stored by MNOs offline for billing and legal purposes, along with their subscription data (e.g., invoice address). This way, instead of MNOs transmit the raw data of this user (according to the pre-defined indicators), this data should be processed by an ϵ-LDP algorithm in the MNOs servers, where ϵ is a pub-

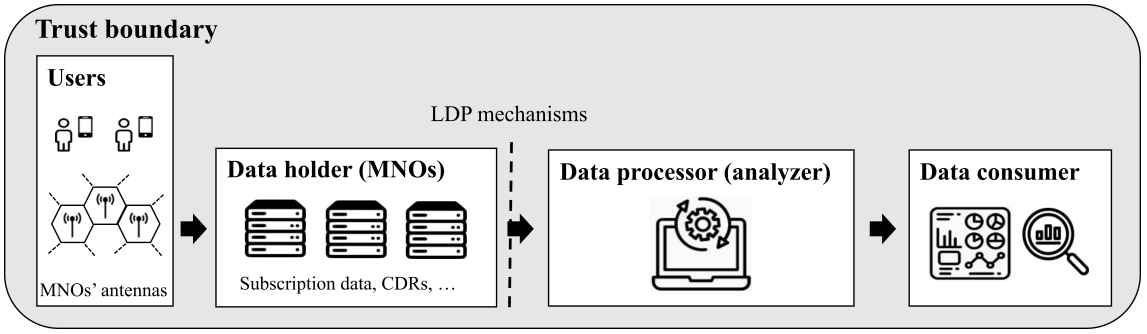

Figure 5.1: Overview of our system model with an LDP-based privacy-preserving solution to sanitize each users' data on-the-fly before transmitting to the analyzer.

lic parameter, before transmitting it to the data processor. **Besides, similar to MNOs CDRs processing systems, the users' IDs should be excluded before transmitting any data to the analyzer. Thus, improving the users' privacy.**

Therefore, the data processor would only store uncorrelated (i.e., no IDs) ϵ-LDP data. At the end of the period of analysis, the analyzer can aggregate these data to produce statistics through multiple frequency estimation, which depends on the LDP algorithm and the public parameter ϵ. Notice that **with our proposal, both users and MNOs are safeguarded as no *raw data* will be collected to analyze human mobility anymore.** *However, ϵ-LDP values must not result in **indirect unique identifiers**.* Indeed, 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.

So, in this chapter, we propose to use the GRR [80] LDP mechanism, which corresponds to the situation where no particular encoding is chosen. In other words, with GRR, ϵ-LDP private reports will become anonymous depending on the size of the attribute (e.g., feminine or masculine, for the gender attribute), thus allowing a longitudinal collection of data. Lastly, we propose to generate mobility reports similar to the FIMU-DB of Section 3.3.1.1, i.e., multidimensional frequency estimates by day and by the union of consecutive days ("*cumulative frequency estimates*").

The remainder of this chapter is organized as follows. In Section 5.2, We extended the analytical analysis of GRR for multidimensional frequency estimates. Next, we explain our proposed LDP-based CDRs processing system in Section 5.3. In Section 5.4, we present our results, its discussions, and review related work. Lastly, in Section 5.5 we present the concluding remarks. The results of Section 5.2 and a preliminary version of the proposed LDP-based CDRs processing system in Section 5.3 with its results were published in a full paper [215] at the 15th IFIP Summer School on Privacy and Identity Management.

5.2/ MULTIDIMENSIONAL FREQUENCY ESTIMATES WITH GRR

In the literature, there are few works for collecting multidimensional data with LDP based on random sampling (i.e., dividing users in groups) [83, 166, 123, 239, 108]. This technique reduces both dimensionality and communication costs, which will also be the focus of this chapter. Let $d \geq 2$ be the total number of attributes, $\mathbf{c} = [c_1, c_2, ..., c_d]$ be the domain size of each attribute, n be the number of users, and ϵ be the whole privacy budget. An intuitive solution is splitting (*Spl*) the privacy budget, i.e., assigning ϵ/d for each attribute. The other solution is based on uniformly sampling without replacement (*Smp*) only r attribute(s) out of d possible ones, i.e., assigning ϵ/r per attribute. Notice that both solutions satisfy ϵ-LDP according to the sequential composition theorem [59]. More visually, Fig. 5.2 illustrates both *Spl* and *Smp* solutions, with $r = 1$ for *Smp*.

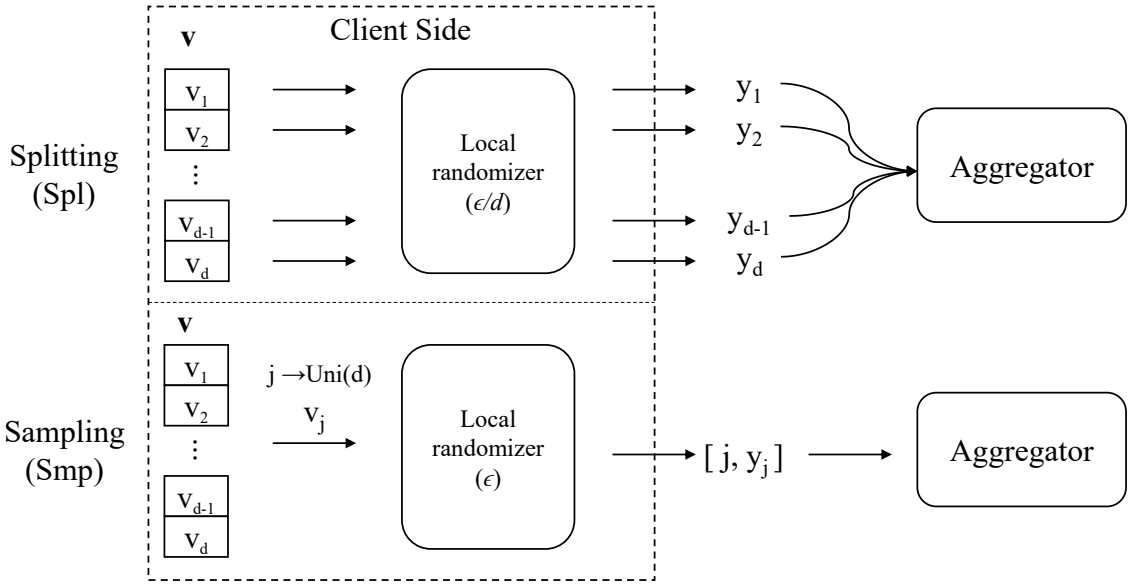

Figure 5.2: State-of-the-art solutions for multidimensional frequency estimates under ϵ-LDP guarantees, where $Uni(d) = Uniform(\{1, 2, ..., d\})$.

For the first case, *Spl*, replacing ϵ by ϵ/d in Eq. 2.8, gives the variance (σ_1^2) of GRR as:

$$\sigma^2_{1,GRR} = \frac{e^{\epsilon/d} + c_j - 2}{n(e^{\epsilon/d} - 1)^2}. \tag{5.1}$$

For the second case, *Smp*, the number of users per attribute is reduced to nr/d. Thus, replacing n by nr/d and ϵ by ϵ/r in Eq. 2.8, gives the variance (σ_2^2) of GRR as:

$$\sigma^2_{2,GRR} = \frac{d(e^{\epsilon/r} + c_j - 2)}{nr(e^{\epsilon/r} - 1)^2}. \tag{5.2}$$

Obviously, if $r = d$ in Eq. (5.2), one has Eq. (5.1). Practically, the objective is reduced to

finding r, which minimizes $\sigma^2_{2,GRR}$. This way, to find the optimal r, we first multiply $\sigma^2_{2,GRR}$ in Eq. (5.2) by ϵ. Without loss of generality, minimizing $\sigma^2_{2,GRR}$ is equivalent to minimizing $\frac{\epsilon e^{\epsilon/r}}{r(e^{\epsilon/r}-1)^2}$. Hence, let $x = r/\epsilon$ be the independent variable, $\sigma^2_{2,GRR}$ can be rewritten as $y = \frac{1}{x} \cdot \frac{e^{1/x}}{(e^{1/x}-1)^2}$ as a function over x. It is not hard to prove that y is an increasing function w.r.t. x and, hence, we have a minimum and optimal when $r = 1$ (a single attribute per user). We highlight that this is a common result in the LDP literature obtained for different protocols and contexts [83, 166, 239, 108, 208, 238, 184, 88].

5.3/ LDP-BASED COLLECTION OF CDRS FOR MOBILITY REPORTS

In this section, according to the system overview in Fig. 5.1, we detail our LDP-based solution (Section 5.3.1) regarding the *Cumulative frequency estimates* scenario outlined in the introduction and its limitations (Section 5.3.2).

5.3.1/ PROPOSED METHODOLOGY

Fig. 5.3 illustrates the overview of our LDP-based CDRs processing system applied to generate mobility reports by days and by the *union* of consecutive 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.

1. **Initialization.** According to the left side of Fig. 5.3, MNOs should define the privacy guarantee ϵ, which is uniform for all users. Let Nb be the whole period of analysis (e.g., total number of days) known a priori, e.g., before an event like the FIMU. So, the data processor should initialize $Nb(Nb+1)/2$ empty databases, *which corresponds to all days and union of consecutive days*. 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.** Similar to MNOs CDRs processing systems (e.g., [53]), MNOs will continue to be responsible for applying a privacy-preserving technique. In our proposal, the privacy-preserving technique corresponds to an LDP-based sanitization model on-the-fly using the GRR [80] protocol explained in Section 2.4. Besides, we assume that MNOs store information about their clients such that each user u_i $(1 \leq i \leq n)$ has a discrete-encoded tuple record $\mathbf{v} = (v_1, v_2, ..., v_d)$, which contains the values of d categorical attributes $A = \{A_1, A_2, ..., A_d\}$ (according to the pre-defined mobility indicators, e.g., as the table in the right side of Fig. 5.3). Since we have multiple attributes, we adopt the *Smp* solution from 5.2,

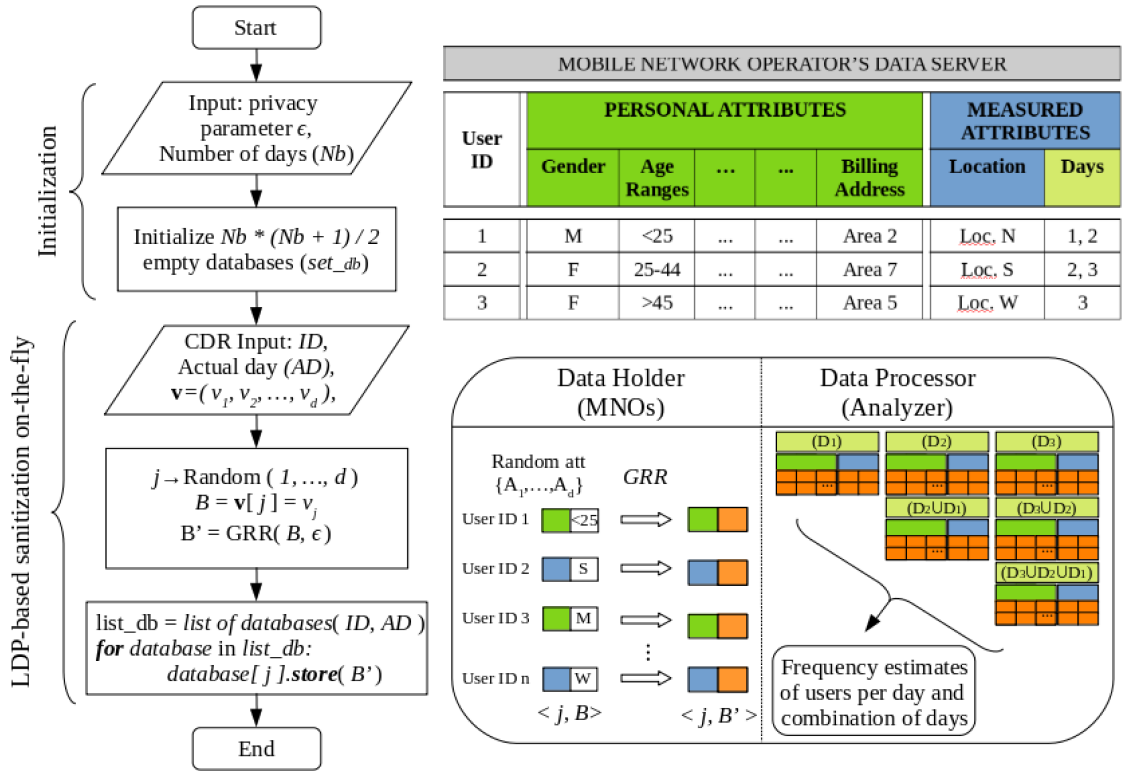

Figure 5.3: Overview of our LDP-based CDRs processing system to generate mobility reports by days and by the *union* of consecutive days.

which randomly samples a **single attribute per user** and uses the whole privacy budget ϵ to sanitize it. For the rest of this chapter, we will refer to this solution as Smp[GRR].

Therefore, we propose that MNOs apply GRR a *single time* (i.e., once and for all) for each users' sampled data $B = v_j$ and consistently use the sanitized value B' for all future reports $\langle j, B' \rangle$. In other words, MNOs would not use any raw data anymore but, rather, an ϵ-LDP version of their clients' data. Since GRR does not utilize any particular encoding, the uncorrelated ϵ-LDP values could be made 'anonymous' within all other reports, thus, allowing longitudinal data collection with no risk of creating a 'unique ID'. Notice that our solution can not ensure "anonymity on-the-fly", but instead, the ϵ-LDP values could probably be "hidden in the crowd" depending on the domain size of the attributes.

Moreover, on the MNOs' side, each CDR contains metadata such as the user's ID and timestamp (Actual Day – AD). Hence, for each user, MNOs calculate a $list_{db}$ that represents which databases (days and union of consecutive days) the ϵ-LDP data should be stored by the data processor (with no ID). For instance, the $list_{db}$ can be calculated by knowing the days this user "was present" (by CDRs) or, similarly, by using Bloom filters [22] to de-duplicate the users' presence throughout days. We

later explain in an example how to calculate $list_{db}$.

3. **Generating statistics.** Throughout the analysis period, the data processor can estimate the frequency of the population for all d attributes for the database of each day and the combinations of past consecutive days. Finally, at the end of the analysis period, the analyst will have $Nb(Nb+1)/2$ databases, with the estimated frequencies for all d attributes in each combination of *union* of consecutive days.

Example to calculate $list_{db}$. To calculate the $list_{db}$ for each user, consider the right side of Fig. 5.3, 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 MNO 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 ϵ-LDP report in the whole analysis. Next, let $AD = 2$. For the same user $ID = 1$, the MNO knows he was present in both two days, hence, his $list_{db} = \{D_2, D_3 \cup D_2\}$ as the previous day his ϵ-LDP report 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 ϵ-LDP report 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 when $AD = 3$.

5.3.2/ LIMITATIONS

The first key limitation we see in our methodology is the storage factor, which is due to collecting users' data per day and union of consecutive days. For instance, data processors 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 the spread of diseases, etc. In addition, there is high power for computation and powerful tools to deal with big data nowadays. One way to smoothen this problem in, e.g., daily scenarios, is to exclude the stored data after retrieving statistics.

Further, similar to the FIMU-DB explained in Chapter 3, 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 consecutive days as it may have very few users per intersection (see our enhanced mobility scenario in Table 4.2 of Chapter 4). 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 developed in Chapter 4, which models our proposed mobility scenario (days and union of consecutive 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 data processors would have to initialize $2^{Nb} - 1$ empty databases (all combinations of intersections of days).

Lastly, the *single time* sanitization step implies always reporting the same sanitized value B' for the unique sampled attribute, which can be effective in the cases where the true client's data does not vary (static) [61, 95]. On the other hand, a measured attribute such as location is dynamic. Therefore, for the users who sample a dynamic attribute, for each different value, a new sanitized value would be generated, thus accumulating the privacy budget ϵ by the sequential composition theorem [59]. Yet, in our privacy-preserving architecture (Fig. 5.1), the collected/stored ϵ-LDP reports are 'uncorrelated' from users, as no ID will be stored. Thus, improving the privacy of users.

5.4/ Results and Discussion

In this section, we present the setup of our experiments in Section 5.4.1. Next, we report the results in Section 5.4.2 obtained by applying our proposed methodology in the MS-FIMU dataset generated in Chapter 4. Lastly, we discuss our work and review related work in Section 5.4.3.

5.4.1/ Setup of Experiments

Environment. All algorithms were implemented in Python 3.8.8 with NumPy 1.19.5 and Numba 0.53.1 libraries. The codes we developed for the preliminary results in paper [215] are available in a Github repository[1]. In all experiments of this manuscript, we report average results over 100 runs as LDP algorithms are randomized.

Dataset. We experimented with the MS-FIMU dataset from Chapter 4. In these experiments, we excluded the data from 'Foreign tourist' users regarding the 'Visitor category' attribute. Hence, the filtered dataset aggregates a population of $87,098$ unique French users with 6 attributes, where 5 are static ('Visitor category' excluded) and 1 is dynamic, along $Nb = 7$ days (on average $\sim 26,000$ unique users per day). For more details about the attributes of this dataset, please refer to Section 4.4.2. Notice that the 'Region' attribute only considers 22 regions in France since we excluded Foreign people.

Evaluation and metrics. Let $Nb = 7$ days be the whole analysis period, we then have $Nb(Nb+1)/2 = 28$ databases considering each day and union of consecutive 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

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

time, we can 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 consecutive days $D_7 \cup D_6 \cup ... \cup D_1$ has all $87,098$ users.

We vary the privacy parameter in the range $\epsilon = [1, 2, 3, 4, 5, 6]$, which is within range of values experimented in the literature for multidimensional data (e.g., in [166] the range is $\epsilon = [0.5, ..., 4]$ and in [239] the range is $\epsilon = [0.1, ..., 10]$).

To evaluate our results, we use the mean squared error (MSE) metric averaged per the number of attributes d. Thus, for each attribute j at time $t \in [1, Nb]$, we compute for each value $v_i \in A_j$ the estimated frequency $\hat{f}(v_i)$ and the real one $f(v_i)$ and calculate their differences. More precisely,

$$MSE_{avg} = \frac{1}{d} \sum_{j \in [1,d]} \frac{1}{|A_j|} \sum_{v_i \in A_j} (f(v_i) - \hat{f}(v_i))^2. \tag{5.3}$$

Methods evaluated. We consider for evaluation the two solutions from Section 5.2:

- Spl[GRR]: Splitting the privacy budget over the number of attributes d, i.e., for each user, send all value with ϵ/d-LDP.
- Smp[GRR]: Sampling a single attribute and send it with the whole privacy budget, i.e., for each user, send a sampled value with ϵ-LDP. This is the solution adopted in our LDP-based CDRs processing system (cf. Fig. 5.3).

5.4.2/ CUMULATIVE FREQUENCY ESTIMATES RESULTS

Fig. 5.4 illustrate for both Spl[GRR] and Smp[GRR] methods, the averaged MSE_{avg} per the number of days Nb (y-axis) according to the privacy parameter ϵ (x-axis). With more details, Fig. 5.5 illustrates for both methods the MSE_{avg} results (y-axis) according to the privacy budget ϵ for each day and the union of consecutive days (x-axis), e.g., '321' refers to $D_3 \cup D_2 \cup D_1$. Lastly, for the sake of illustration, Fig. 5.6 illustrates multidimensional frequency estimates for a single day (D_7) and for the union of all consecutive days ($D_7 \cup D_6 \cup ... \cup D_1$) using the adopted Smp[GRR] solution and $\epsilon = 1$.

As one can notice in Fig. 5.4, overall, the proposed Smp[GRR] solution adopted in our LDP-based CDRs processing system consistently and considerably outperforms the baseline Spl[GRR]. In Fig. 5.5, except for $\epsilon = 1$, the curves of Smp[GRR] are under even to the best one of Spl[GRR] using the highest privacy budget $\epsilon = 6$. As also highlighted in the literature [166, 83, 108], privacy budget splitting is sub-optimal, which leads to higher estimation error. Indeed, in a multidimensional setting, the combination of privacy budget splitting and high numbers of values in a given attribute (e.g., *Region* with 22 values) leads

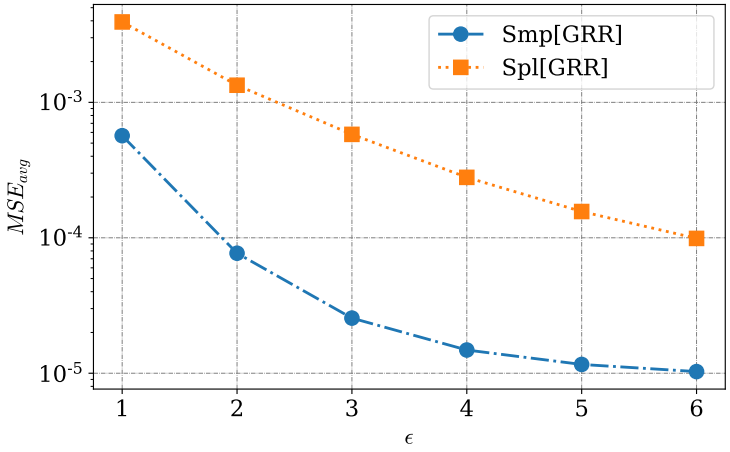

Figure 5.4: Averaged MSE_{avg} per the number of days Nb (y-axis) varying ϵ (x-axis) on the MS-FIMU dataset comparing Spl[GRR] and Smp[GRR].

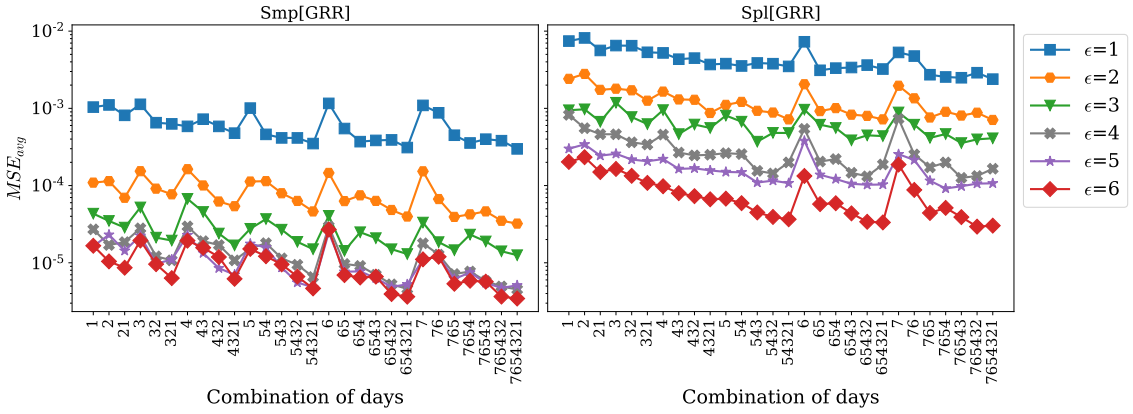

Figure 5.5: MSE_{avg} (y-axis) analysis comparing Smp[GRR] (left-side plot) and Spl[GRR] (right-side plot) by varying the privacy budget ϵ on each combination of days (x-axis) individually.

to lower data utility even for high privacy regimes. On the other hand, the Smp[GRR] 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 due to observing a sample instead of the entire population.

Moreover, in Figs. 5.5 and 5.6, it is noteworthy that the MSE_{avg} decreases as the data size increases. Intuitively, this is due to LDP, which requires a large amount of data to guarantee a good balance of noise. In our case, single days (e.g., D_7) have less users comparing to the union of all consecutive days (e.g., $D_7 \cup D_6 \cup ... \cup D_1$) and, hence, single days are generally the peak-values in Fig. 5.5. Indeed, these results are consistent with Eqs. (5.1) and (5.2), where the variances are decreasing functions over the number of users n. Yet, these peak values are smoothed using Smp[GRR], which induces less error by sampling a single attribute for each user.

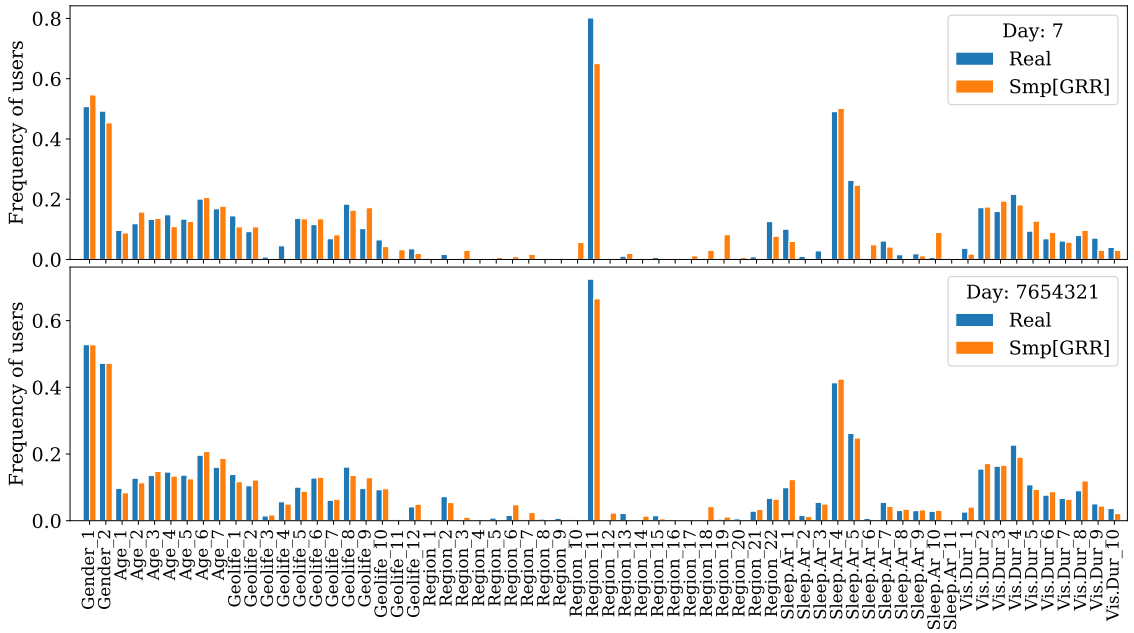

Figure 5.6: Comparison between real and estimated frequencies for a single day (D_7) and to the union of all consecutive days ($D_7 \cup D_6 \cup ... \cup D_1$) using the adopted Smp[GRR] solution and $\epsilon = 1$.

Lastly, we highlight that the objective of our experiments was to measure the accuracy loss (based on the MSE_{avg} error metric) of using our LDP-based "sanitization on-the-fly" system in comparison with the original statistics produced by an "anonymity on-the-fly" based system. As shown in the results, accurate multidimensional frequency estimates could be achieved for practical purposes with strong privacy guarantees (see, e.g., Fig. 5.6 with $\epsilon = 1$). On the other hand, in terms of the overall privacy budget ϵ per user, in the worst case, the sequential composition theorem [59] applies for each data release. As also pointed out in [232, Section 8.4, Table 2] and in [219], real-world DP systems utilize ϵ as large as the ones experimented in this manuscript on daily basis. Thus, some future implementation of our LDP-based CDRs processing system to generate mobility reports is a potential perspective.

5.4.3/ Discussion and Related Work

As reviewed in Chapters 1 and 2, mobile phone CDRs have been largely used to analyze human mobility in several contexts, e.g., the spread of infectious diseases [156, 44, 65, 218, 186, 237], natural disasters [127, 182, 44], tourism [53, 62, 194], and so on. However, concerning privacy, de Montjoye et al. [49] show that humans follow particular patterns, which allows predicting human mobility with high accuracy. For instance, in a dataset of 1.5 million users, the authors showed that 95% of this population can be re-identified using four approximate locations and their timestamps. Besides, Zang and

Bolot [38] have performed extensive experiments showing that the anonymization of location data from CDRs using *k*-anonymity [18, 20] leads to privacy risks. Further, in non-technical papers, de Montjoye et al. [122] discuss the conscientious use of mobile phone data for mobility analytics, and Buckee [56] highlights both the importance of collecting CDRs to analyzing human mobility in low-income countries and the privacy concerns that rise up.

Because of these privacy issues, MNOs tend to publish aggregated mobility data [218, 109, 237, 135, 53], e.g., the number of users by coarse location at a given timestamp or the number of users in a single location (cf. Section 3.3.1.1). However, as recent studies have shown, even aggregated mobility data can be subject to membership inference attacks [103, 198] and users' trajectory recovery attack [135, 109]. More precisely, the later authors in [135, 109] showed that their attack reaches accuracies as high as $73\% \sim 91\%$, suggesting generalization and perturbation through DP [27, 26, 59] as a means to mitigate this attack. Therefore, it is vital to deploying systems that allow analyzing human mobility (e.g., through CDRs) with strong privacy-preserving guarantees.

With these elements in mind and with the motivating questions $\mathbf{Q}_1$ and $\mathbf{Q}_2$ from the beginning of this chapter, we have proposed a solution beyond "anonymity on-the-fly" since aggregated location data are still at risk of leaking private information. Indeed, our solution considers "sanitization on-the-fly" with an LDP protocol, in which rather than transmitting aggregated raw data for the analyzer, we propose that MNOs sanitize each users' data independently (as if it was made by the user) and send it to the *untrusted analyzer*.

As we present in this chapter, implementing the Smp[GRR] solution in our methodology could ensure that ϵ-LDP private reports will not become indirect *unique IDs*. The reason behind this is because no particular encoding is used with GRR and, thus, ϵ-LDP values are generic to any user. So, it is possible to utilize the sanitized value in longitudinal studies if the domain size of attributes is not big. Besides, notice that each time users connect, MNOs will always report the same 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), MNOs will never report the remaining $d - 1$ attributes, which were not sampled. Lastly, our solution would also safeguard MNOs as no *raw data* would be shared with the analyzer for the purpose of human mobility analysis, but, rather, ϵ-LDP values that are robust to post-processing. One clear limitation of our LDP-based CDRs processing system is that the recent privacy amplification by shuffling [141, 149, 184, 202, 224] does not apply. Although all users' IDs are excluded, the signals' order is not hidden due to "sanitization on-the-fly". That is, the ϵ-LDP reports are not aggregated in "batches" to provide some "anonymity" and profit from amplification. Therefore, extending our solution to the shuffle DP model is a potential and intended perspective.

5.5/ CONCLUSION

This chapter investigated the problem of collecting and analyzing CDRs-based data to generate multidimensional frequency estimates throughout time. We proposed an LDP-based CDRs processing system as an extension of "anonymity on-the-fly" to satisfy "*sanitization* on-the-fly", thus, providing higher privacy guarantees for each user. With our proposal, we can have preliminary answers to the motivating questions $\mathbf{Q}_1$ and $\mathbf{Q}_2$ highlighted at the beginning of this chapter. That is, such a privacy-preserving system would allow MNOs to share the sanitized data with *untrusted analyzers*, with a more strict setting that allows sanitizing each data independently on-the-fly. As shown in the results, the proposed LDP-based CDRs processing system using Smp[GRR] achieves accurate multidimensional frequency estimates for practical purposes (*cf.* Fig. 5.6, for example), proving its effectiveness in producing mobility reports as the original ones from OBS.

On the one hand, this is because GRR has low utility loss for attributes with small domain sizes. On the other hand, if MNOs intend to pre-define a mobility indicator on a higher domain (e.g., the number of people in each $\sim 1,000$ bus stops of a given city), other protocols like OUE [108] could provide higher data utility, as its variance does not depend on the domain size. However, since OUE is based on unary-encoding (cf. Section 2.4), it would probably generate a sanitized value similar to a *unique ID*. In other words, analyzers would be able to use the *unique* OUE-based reports to track individuals across many days. One possible solution would be using two rounds of sanitization (i.e., *memoization* [61, 95]), also mentioned in Section 2.4. Indeed, this is one of the core contributions of the next Chapter 6, which investigates how to improve the utility of LDP protocols for longitudinal (based on memoization) and multidimensional frequency estimates.

6

Multidimensional Frequency Estimates Over Time With LDP: Utility Focus

In Chapter 5, we focused on a more **practical** perspective for the problem of generating multidimensional mobility reports throughout time from CDRs. In this chapter, we abstracted this problem and, thus, we contribute on the **theoretical** aspect by optimizing **the utility** of LDP protocols for *longitudinal* and *multidimensional* frequency estimates. This way, the more the estimated frequencies approximate the real ones, the more ML models can take advantage of when performing learning/prediction tasks [183]. Notice that **our solutions are generic to any LDP application scenario** (e.g., collecting user behavior in software [61, 95, 106]). We invite the reader to refer to Chapter 2 for the background on LDP.

6.1/ Introduction

In this chapter, we focus on the problem of private frequency (or histogram) estimation of multiple attributes throughout time with LDP. As in previous Chapter 5, we assume there are d attributes $A = \{A_1, A_2, ..., A_d\}$, where each attribute A_j with a discrete domain has a specific number of values $c_j = |A_j|$. Each user u_i for $i \in \{1, 2, ..., n\}$ has a tuple $\mathbf{v}^{(i)} = (v_1^{(i)}, v_2^{(i)}, ..., v_d^{(i)})$, where $v_j^{(i)}$ represents the value of attribute A_j in record $\mathbf{v}^{(i)}$. Thus, for each attribute A_j at time $t \in [1, \tau]$, the aggregator's goal is to estimate a c_j-bins histogram, including the frequency of all values in A_j.

On tackling both longitudinal and multidimensional settings, one needs to consider the allocation of the privacy budget, which can grow extremely quickly due to the composition theorem [59]. So, first, we focus on solving the multidimensional aspect with a random sampling-based solution [83, 166, 123, 239], also used in Chapter 5. Next, we considered

the *memoization*-based framework [61, 95, 184] to solve the longitudinal setting, which allows having an upper bound to the privacy budget. In both cases, we extended the analysis of three state-of-the-art protocols, namely, GRR [80], OUE [108], and SUE [61], presented in Section 2.4. Thus, combining the optimal cases of each setting, we propose a new solution named Adaptive LDP for LOngitudinal and Multidimensional FREquency Estimates (ALLOMFREE). We demonstrate through experimental validations using four real-world datasets the advantages of ALLOMFREE over state-of-the-art protocols [61, 108], with a gain of accuracy, on average, ranging from 10% up to 55% with the analyzed range of ϵ-LDP guarantees.

The rest of this chapter is organized as follows. In Section 6.2, we extend the analysis of OUE and SUE to multidimensional data collections. In Section 6.3 we present the *memoization*-based framework for longitudinal data collections, the extension and analysis of longitudinal GRR and longitudinal UE-based protocols; the numerical evaluation of their performance, and we present our ALLOMFREE solution. In Section 6.4, we present experimental results, discuss our results and review related work. Lastly, in Section 6.5, we present the concluding remarks. The development in Sections 6.2 and 6.3 and the results presented in Section 6.4 were submitted as part of a full article [214] to the Digital Communications and Networks journal.

6.2/ MULTIDIMENSIONAL FREQUENCY ESTIMATES WITH LDP

As reviewed in Section 5.2, there are mainly two solutions for collecting multidimensional data with LDP (see Fig. 5.2). In this section, we will follow the same development used in Section 5.2 for two other protocols, namely, SUE and OUE. Let $d \geq 2$ be the total number of attributes, $\mathbf{c} = [c_1, c_2, ..., c_d]$ be the domain size of each attribute, n be the number of users, and ϵ be the whole privacy budget.

For the first case, *Spl*, replacing ϵ by ϵ/d in Eqs. (2.10) and (2.11) give the variances (σ_1^2) of SUE and OUE, respectively, as:

$$\begin{aligned} \sigma^2_{1,SUE} &= \frac{e^{\epsilon/2d}}{n(e^{\epsilon/2d}-1)^2}, \\ \sigma^2_{1,OUE} &= \frac{4e^{\epsilon/d}}{n(e^{\epsilon/d}-1)^2}. \end{aligned} \tag{6.1}$$

For the second case, *Smp*, the number of users per attribute is reduced to nr/d. Thus, replacing n by nr/d and ϵ by ϵ/r in Eqs. 2.10 and 2.11 give the variances (σ_2^2) of SUE and OUE, respectively, as:

$$\sigma^2_{2,SUE} = \frac{d(e^{\epsilon/2r})}{nr(e^{\epsilon/2r}-1)^2},$$
$$\sigma^2_{2,OUE} = \frac{d(4e^{\epsilon/r})}{nr(e^{\epsilon/r}-1)^2}. \tag{6.2}$$

Obviously, if $r = d$ in Eq. (6.2), one has Eq. (6.1). Practically, the objective is reduced to finding r, which minimizes σ_2^2 for each protocol. This way, to find the optimal r for each protocol, we first multiply each σ_2^2 in Eq. (6.2) by ϵ. Without loss of generality, minimizing $\sigma^2_{2,SUE}$ and $\sigma^2_{2,OUE}$ is equivalent to minimizing $\frac{\epsilon e^{\epsilon/2r}}{r(e^{\epsilon/2r}-1)^2}$ and $\frac{\epsilon e^{\epsilon/r}}{r(e^{\epsilon/r}-1)^2}$ (similar to GRR in Section 5.2), respectively. Hence, let $x = r/\epsilon$ be the independent variable, $\sigma^2_{2,OUE}$ can be rewritten as $y_1 = \frac{1}{x} \cdot \frac{e^{1/x}}{(e^{1/x}-1)^2}$ and $\sigma^2_{2,SUE}$ can be rewritten as $y_2 = \frac{1}{x} \cdot \frac{e^{1/2x}}{(e^{1/2x}-1)^2}$ as functions over x. It is not hard to prove that both y_1 and y_2 are increasing functions w.r.t. x and, hence, we have a minimum and optimal when $r = 1$ (a single attribute per user) for both protocols too.

Therefore, in this chapter, we adopt the multidimensional setting *Smp* with $r = 1$. In this setting, users tell the data collector which attribute was sampled, and its perturbed value ensuring ϵ-LDP by applying either GRR or UE-based protocols; the data analyst server would not receive any information about the remaining $d - 1$ attributes. .

6.3/ Longitudinal Frequency Estimates with LDP

In this section, we present the *memoization*-based framework for longitudinal data collections (Section 6.3.1). Next, we present the analysis of longitudinal GRR (Section 6.3.2) and longitudinal UE-based protocols (Section 6.3.3). Lastly, we evaluate numerically the extended longitudinal protocols (Section 6.3.4) and we propose our ALLOMFREE solution (Section 6.3.5).

6.3.1/ Memoization-based data collection with LDP

In the literature, many works study how to collect and analyze categorical data longitudinally based on *memoization* [61, 95, 184]. The key idea behind memoization is using two sanitization processes. The first round (RR_1) replaces the real value B with a sanitized one B' with a higher epsilon (ϵ_∞). Whenever one intends to report B, B' shall be reused to produce other sanitized versions B'' with lower epsilon values. Notice that the second sanitization (RR_2) is a *must* to avoid 'averaging attacks', in which adversaries can reconstruct the true value from multiple sanitized versions of it. This technique allows achieving privacy over time with an upper bound value of ϵ_∞-LDP.

Let $A_j = \{v_1, v_2, ..., v_{c_j}\}$ be a set of $c_j = |A_j|$ values of a given attribute and let ϵ be the privacy budget. In this chapter, for both RR_1 and RR_2 steps, we will apply either GRR, SUE, or OUE. The unbiased estimator in Eq. (2.4) for the frequency $f(v_i)$ of each value v_i for $i \in [1, c_j]$ is now extended to:

$$\hat{f}_L(v_i) = \frac{N_i - nq_1(p_2 - q_2) - nq_2}{n(p_1 - q_1)(p_2 - q_2)}, \tag{6.3}$$

in which N_i is the number of times the value v_i has been reported, n is the total number of users, p_1 and q_1 are the parameters used by an LDP protocol for RR_1, and p_2 and q_2 are the parameters used by an LDP protocol for RR_2.

Theorem 2. *The estimation result $\hat{f}_L(v_i)$ in Eq. (6.3) is an unbiased estimation of $f(v_i)$ for any value $v_i \in A_j$.*

Proof.

$$\begin{aligned} E[\hat{f}_L(v_i)] &= E\left[\frac{N_i - nq_1(p_2 - q_2) + nq_2}{n(p_1 - q_1)(p_2 - q_2)}\right] \\ &= \frac{E[Ni]}{n(p_1 - q_1)(p_2 - q_2)} - \frac{q_1(p_2 - q_2) - q_2}{(p_1 - q_1)(p_2 - q_2)}. \end{aligned}$$

Let us focus on

$$\begin{aligned} E[N_i] = nf(v_i)\,(p_1p_2 + q_2\,(1 - p_1)) \\ + n\,(1 - f(v_i))\,(p_2q_1 + q_2\,(1 - q_1))\,. \end{aligned}$$

Thus,

$$E[\hat{f}_L(v_i)] = f(v_i).$$

□

Theorem 3. *The variance of the estimation in Eq. (6.3) is:*

$$\begin{aligned} Var(\hat{f}_L(v_i)) = \frac{\gamma(1-\gamma)}{n(p_1 - q_1)^2(p_2 - q_2)^2}, \text{ where} \\ \gamma = f(v_i)\,(2p_1p_2 - 2p_1q_2 + 2q_2 - 1) + p_2q_1 + q_2(1 - q_1). \end{aligned} \tag{6.4}$$

Proof. Thanks to Eq. (6.3) we have

$$Var(\hat{f}_L(v_i)) = \frac{Var(N_i)}{n^2(p_1 - q_1)^2(p_2 - q_2)^2}.$$

Since N_i is the number of times the value v_i is observed, it can be defined as $N_i = \sum_{z=1}^{n} X_z$

where X_z is equal to 1 if the user z, $1 \leq z \leq n$ reports value v_i, and 0 otherwise. We thus have $Var(N_i) = \sum_{z=1}^{n} Var(X_z) = nVar(X)$. Since all the users are independent,

$$P(X = 1) = P(X^2 = 1) = f(v_i)(2p_1p_2 - 2p_1q_2 + 2q_2 - 1) + p_2q_1 + q_2(1 - q_1) = \gamma.$$

We thus have $Var(X) = \gamma - \gamma^2 = \gamma(1 - \gamma)$ and, finally,

$$Var(\hat{f}_L(v_i)) = \frac{\gamma(1-\gamma)}{n(p_1 - q_1)^2(p_2 - q_2)^2}.$$

□

In this chapter, we will use the *approximate variance*, in which $f(v_i) = 0$ in Eq. (6.4), which gives:

$$Var^*(\hat{f}_L(v_i)) = \frac{(p_2q_1 - q_2(q_1 - 1))(-p_2q_1 + q_2(q_1 - 1) + 1)}{n(p_1 - q_1)^2(p_2 - q_2)^2}. \tag{6.5}$$

6.3.2/ LONGITUDINAL GRR (L-GRR): DEFINITION AND ϵ-LDP STUDY

Let $V = \{v_1, v_2, ..., v_{c_j}\}$ be a set of c_j values of a given attribute and let v_i be the real value. We now describe an extension of GRR for longitudinal studies; we refer to this protocol as L-GRR for the rest of this chapter. First, GRR does not require any particular encoding (direct encoding [108]). Next, there are two rounds of sanitization, RR_1 and RR_2 applying GRR, described in the following.

1. $RR_1[GRR]$: Memoize a value B' such that

$$B' = \begin{cases} v_i, & \text{with probability } p_1, \\ v_{k \neq v_i}, & \text{with probability } q_1 = \frac{1-p_1}{c_j - 1}, \end{cases}$$

in which p_1 and q_1 control the level of longitudinal ϵ_∞-LDP. The value B' shall be reused as the basis for all future reports on the real value v_i.

2. $RR_2[GRR]$: Generate a reporting B'' such that

$$B'' = \begin{cases} B', & \text{with probability } p_2, \\ v_{k \neq B'}, & \text{with probability } q_2 = \frac{1-p_2}{c_j - 1}, \end{cases}$$

in which B'' is the report to be sent to the server.

Visually, Fig. 6.1 illustrates the probability tree of the L-GRR protocol. In the first round of

sanitization, RR_1, our proposed L-GRR applies GRR with $p_1 = Pr[B' = v_i|B = v_i] = \frac{e^{\epsilon_\infty}}{e^{\epsilon_\infty}+c_j-1}$ and $q_1 = Pr[B' = v_i|B = v_{k\neq i}] = \frac{1-p_1}{c_j-1} = \frac{1}{e^{\epsilon_\infty}+c_j-1}$ (highlighted in the middle of Fig. 6.1), where $c_j = |A_j|$. As discussed in Section 2.4.2, this *permanent* memoization satisfies ϵ_∞-LDP since $\frac{p_1}{q_1} = e^{\epsilon_\infty}$, which is the upper bound.

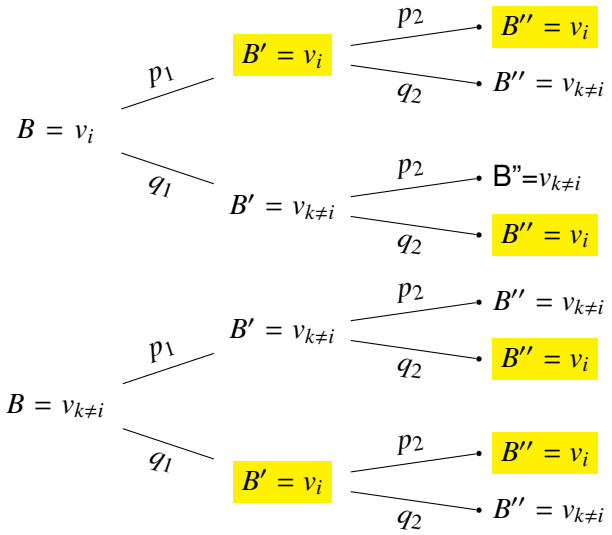

Figure 6.1: Probability tree for two rounds of sanitization using GRR (L-GRR).

On the other hand, with a single collection of data, the attacker's knowledge of v_i comes only from B'', which is generated using two randomization steps with GRR. This provides a higher level of privacy protection [61]. From Fig. 6.1, we can obtain the following conditional probabilities:

$$\Pr[B''|B] = \begin{cases} \Pr[B'' = v_i|B = v_i] = p_1p_2 + q_1q_2 \\ \Pr[B'' = v_{k\neq i}|B = v_i] = p_1q_2 + q_1p_2 \\ \Pr[B'' = v_i|B = v_{k\neq i}] = p_1q_2 + q_1p_2 \\ \Pr[B'' = v_{k\neq i}|B = v_{k\neq i}] = p_1p_2 + q_1q_2 \end{cases}$$

Let $p_s = \Pr[B'' = v_i|B = v_i]$ and $q_s = \Pr[B'' = v_i|B = v_{k\neq i}]$ (highlighted in far right of Fig. 6.1), with the second round of sanitization, $RR_2[GRR]$, our proposed L-GRR protocol satisfies ϵ_1-LDP since $\frac{p_s}{q_s} = e^{\epsilon_1}$. Notice that ϵ_1 corresponds to a single report (lower bound) and its extension to infinity reports is limited by ϵ_∞ (upper bound) since $RR_2[GRR]$ uses as input the output of $RR_1[GRR]$. More specifically, the calculus of ϵ_1 for L-GRR is:

$$\epsilon_1 = \ln\left(\frac{p_1p_2 + q_1q_2}{p_1q_2 + q_1p_2}\right) \tag{6.6}$$

in which $p_1 = \frac{e^{\epsilon_\infty}}{e^{\epsilon_\infty}+c_j-1}$, $q_1 = \frac{1-p_1}{c_j-1}$, and both p_2 and q_2 are selectable according with ϵ_∞, ϵ_1,

and c_j, calculated as:

$$p_2 = \frac{e^{\epsilon_1+\epsilon_\infty} - 1}{-c_j e^{\epsilon_1} + \left(c_j - 1\right) e^{\epsilon_\infty} + e^{\epsilon_1} + e^{\epsilon_1+\epsilon_\infty} - 1}, \qquad q_2 = \frac{1 - p_2}{c_j - 1}. \tag{6.7}$$

The estimated frequency $\hat{f}_L(v_i)$ that a value v_i occurs for $i \in [1, c_j]$ is calculated using Eq. (6.3). Lastly, one can calculate the L-GRR approximate variance by replacing the resulting p_1, q_1, p_2, q_2 parameters into Eq. (6.5).

6.3.3/ LONGITUDINAL UE (L-UE): DEFINITION AND ϵ-LDP STUDY

We now describe UE-based protocols for longitudinal studies; we refer to this protocol as L-UE for the rest of this chapter. Let $V = \{v_1, v_2, ..., v_{c_j}\}$ be a set of c_j values of a given attribute and let v_i be the real value. First, $Encode(v) = B$ (unary encoding), where $B = [0, 0, ..., 1, 0, ...0]$, a c_j-bit array where only the v-th position is set to one. Next, there are two rounds of sanitization, RR_1 and RR_2 applying UE-based protocols, described in the following.

1. $RR_1[UE]$: For each bit i, $1 \leq i \leq c_j$ in B, memoize a value B' such that

$$P(B'_i = 1) = \begin{cases} p_1, & \text{if } B_i = 1 \text{ and} \\ q_1, & \text{if } B_i = 0, \end{cases}$$

   in which p_1 and q_1 control the level of longitudinal ϵ_∞-LDP. The value B' shall be reused as the basis for all future reports on the real value v_i.

2. $RR_2[UE]$: For each bit i, $1 \leq i \leq c_j$ in B', generate a reporting B'' that

$$P(B''_i = 1) = \begin{cases} p_2, & \text{if } B'_i = 1 \text{ and} \\ q_2, & \text{if } B'_i = 0, \end{cases}$$

   in which B'' is the report to be sent to the server.

Visually, Fig. 6.2 illustrates the probability tree of the L-UE protocol. **One natural question emerges: how to select the parameters $\{p_1, q_1, p_2, q_2\}$ in order to optimize the utility of this L-UE protocol?** One can see $RR_1[UE]$ as a *permanent* sanitization and $RR_2[UE]$ as a 'small' perturbation to avoid averaging attacks and keep privacy over time.

Based on SUE and OUE, we are then left with four options: two known solutions that strictly use only OUE or SUE parameters in both sanitization steps and two proposed

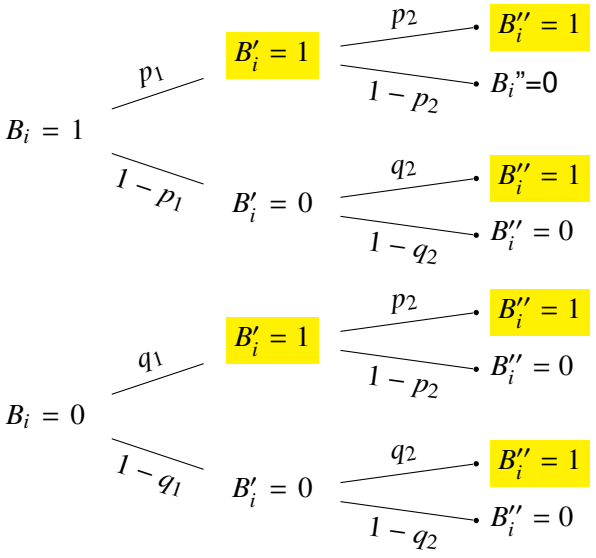

Figure 6.2: Probability tree for two rounds of sanitization using UE (L-UE).

settings that combine both OUE and SUE. These four L-UE protocols are summarized below:

I both sanitization steps with OUE (L-OUE);

II both sanitization steps with SUE (L-SUE);

III starting with OUE and then with SUE (L-OSUE);

IV starting with SUE and then with OUE (L-SOUE);

in which, L-SUE is the well-known Basic-RAPPOR protocol [61], L-OUE is the state-of-the-art OUE protocol [108] with memoization, and both L-OSUE and L-SOUE are proposed in this chapter.

As presented in [108], the OUE variance in Eq. (2.11) is smaller than the SUE variance in Eq. (2.10) and, therefore, the former can provide higher utility than the latter for RR_1. On the other hand, we argue that OUE might be too strict for RR_2 since the parameter $p_2 = 1/2$ is constant. Thus, we hypothesize that option III (i.e., L-OSUE) is the most suitable one. Without loss of generality, **the following analyses are done only for L-OSUE**, which can be easily extended to any of the other combinations.

In the first round of sanitization, RR_1, our solution L-OSUE applies OUE with $p_1 = Pr[B'_i = 1|B_i = 1] = \frac{1}{2}$ and $q_1 = Pr[B'_i = 1|B_i = 0] = \frac{1}{e^{\epsilon_\infty}+1}$ (highlighted in the middle of Fig. 6.2). As discussed in Section 2.4.3, this *permanent* memoization satisfies ϵ_∞-LDP since $\frac{p_1(1-q_1)}{(1-p_1)q_1} = e^{\epsilon_\infty}$, which is the upper bound.

Following the same development as for L-GRR, on the other hand, with a single collection of data, the attacker's knowledge of $B = Encode(v)$ comes only from B'', which is

generated using two randomization steps with OUE and SUE, respectively. This provides a higher level of privacy protection [61]. From Fig. 6.2, we can obtain the following conditional probabilities according to each bit $i \in [1, c_j]$:

$$\Pr[B_i''|B_i] = \begin{cases} \Pr[B_i'' = 1|B_i = 1] = p_1p_2 + (1-p_1)q_2 \\ \Pr[B_i'' = 0|B_i = 1] = p_1(1-p_2) + (1-p_1)(1-q_2) \\ \Pr[B_i'' = 1|B_i = 0] = q_1p_2 + (1-q_1)q_2 \\ \Pr[B_i'' = 0|B_i = 0] = q_1(1-p_2) + (1-q_1)(1-q_2) \end{cases}$$

Let $p_s = \Pr[B_i'' = 1|B_i = 1]$ and $q_s = \Pr[B_i'' = 1|B_i = 0]$ (highlighted in far right of Fig. 6.2), with the second round of sanitization, $RR_2[SUE]$, our proposed L-OSUE protocol satisfies ϵ_1-LDP since $\frac{p_s(1-q_s)}{(1-p_s)q_s} = e^{\epsilon_1}$. Notice that ϵ_1 corresponds to a single report (lower bound) and its extension to infinity reports is limited by ϵ_∞ (upper bound) since $RR_2[SUE]$ uses as input the output of $RR_1[OUE]$. More specifically, the calculus of ϵ_1 for L-OSUE (or L-UE protocols in general) is:

$$\epsilon_1 = \ln\left(\frac{(p_1p_2 - q_2(p_1-1))(p_2q_1 - q_2(q_1-1)-1)}{(p_2q_1 - q_2(q_1-1))(p_1p_2 - q_2(p_1-1)-1)}\right), \tag{6.8}$$

in which, for L-OSUE, we have $p_1 = \frac{1}{2}$, $q_1 = \frac{1}{e^{\epsilon_\infty}+1}$, and both p_2 and q_2 are symmetric ($p_2 + q_2 = 1$) and selectable according to ϵ_∞ and ϵ_1, calculated as:

$$\begin{aligned} p_2 &= \frac{1 - e^{\epsilon_1+\epsilon_\infty}}{e^{\epsilon_1} - e^{\epsilon_\infty} - e^{\epsilon_1+\epsilon_\infty} + 1}, \\ q_2 &= 1 - p_2. \end{aligned} \tag{6.9}$$

Similarly, the estimated frequency $\hat{f}_L(v_i)$ that a value v_i occurs for $i \in [1, c_j]$ is calculated using Eq. (6.3). Lastly, one can calculate the L-OSUE (or L-UE protocols in general) approximate variance by replacing the resulting p_1, q_1, p_2, q_2 parameters into Eq. (6.5).

6.3.4/ Numerical evaluation of L-GRR and L-UE protocols

In this subsection, we evaluate numerically the approximate variance of all developed longitudinal protocols, namely, L-GRR and the four UE-based options namely L-OUE, L-SUE, L-OSUE, and L-SOUE, respectively. As aforementioned, once defined both ϵ_∞ and ϵ_1 privacy guarantees, one can obtain the parameters p_1 and q_1 depending on ϵ_∞, and the parameters p_2 and q_2 depending on both ϵ_∞ and ϵ_1 (and the domain size c_j for L-GRR) as given in Eq. (6.7) for L-GRR and in Eq. (6.9) for L-OSUE.

Next, once computed the parameters $\{p_1, q_1, p_2, q_2\}$, one can calculate the approximate

variance with Eq. (6.5) for each protocol. In other words, following our proposal, one has to set both the upper (ϵ_∞) and lower (ϵ_1) bounds of the privacy guarantees. For example, let $\epsilon_\infty = 2$, one might want that the first ϵ_1-LDP report to have high privacy such as $\epsilon_1 = 0.1$, i.e., $\epsilon_1 = 0.05\epsilon_\infty$ (**we will use this percentage notation to set up the privacy guarantees**).

Table 6.1 exhibits numerical values of the approximate variance using Eq. (6.5) for all longitudinal protocols with $n = 10000$, $\epsilon_\infty = [0.5, 1.0, 2.0, 4.0]$ (as in [108]), and $\epsilon_1 = \{0.6\epsilon_\infty, 0.5\epsilon_\infty, 0.4\epsilon_\infty, 0.3\epsilon_\infty, 0.2\epsilon_\infty, 0.1\epsilon_\infty\}$. For values of ϵ_1 higher than $0.6\epsilon_\infty$, neither L-OUE nor L-SOUE could satisfy some values of ϵ_1 because of the constant $p_2 = 1/2$ in RR_2. Yet, it is not desirable to have higher values of ϵ_1 and, thus, we did not consider values above $0.6\epsilon_\infty$ in our analysis. Besides, Table 6.2 exhibits numerical values for non-longitudinal GRR, OUE, and SUE protocols, which allows evaluating how utility degrades with a second step of sanitization.

Table 6.1: Numerical values of Eq. (6.5) (i.e., $Var^*[\hat{f}_L(v_i)]$) for L-GRR and L-UE protocols with different ϵ_∞ and ϵ_1 privacy guarantees, following $\epsilon_1 = \{0.6\epsilon_\infty, 0.5\epsilon_\infty, 0.4\epsilon_\infty, 0.3\epsilon_\infty, 0.2\epsilon_\infty, 0.1\epsilon_\infty\}$, respectively.

ϵ_1	Privacy Guarantees	L-GRR			L-UE			
		$c_j = 2$	$c_j = 32$	$c_j = 2^{10}$	L-OSUE	L-SUE	L-SOUE	L-OUE
$0.6\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.30$	0.001103	0.980969	26706	0.004411	0.004436	0.005306	0.005549
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.60$	0.000270	0.125036	3153	0.001078	0.001103	0.001234	0.001347
	$\epsilon_\infty = 2.0, \epsilon_1 = 1.20$	0.000062	0.006327	117	0.000247	0.000270	0.000264	0.000310
	$\epsilon_\infty = 4.0, \epsilon_1 = 2.40$	0.000011	0.000078	0.25903	0.000044	0.000062	0.000045	0.000057
$0.5\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.25$	0.001592	2.088372	60218	0.006367	0.006392	0.007336	0.007611
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.50$	**0.000392**	0.268074	7198	**0.001567**	**0.001592**	0.001740	0.001872
	$\epsilon_\infty = 2.0, \epsilon_1 = 1.00$	**0.000092**	0.013926	281	**0.000368**	**0.000392**	0.000389	0.000447
	$\epsilon_\infty = 4.0, \epsilon_1 = 2.00$	**0.000018**	0.000188	0.74088	**0.000072**	**0.000092**	0.000073	0.000092
$0.4\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.20$	0.002492	4.530779	135874	0.009967	0.009992	0.011012	0.011324
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.40$	0.000617	0.586823	16443	0.002467	0.002492	0.002658	0.002812
	$\epsilon_\infty = 2.0, \epsilon_1 = 0.80$	0.000148	0.031552	673	0.000593	0.000617	0.000617	0.000690
	$\epsilon_\infty = 4.0, \epsilon_1 = 1.60$	0.000032	0.000484	2.12772	0.000127	0.000148	0.000128	0.000156
$0.3\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.15$	0.004436	10	329836	0.017744	0.017769	0.018863	0.019214
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.30$	0.001103	1.398568	40412	0.004411	0.004436	0.004620	0.004799
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.60$	0.000270	0.078202	1737	0.001078	0.001103	0.001106	0.001198
	$\epsilon_\infty = 2.0, \epsilon_1 = 1.20$	0.000062	0.001389	6	0.000247	0.000270	0.000248	0.000291
$0.2\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.10$	0.009992	30	972656	0.039967	0.039992	0.041148	0.041536
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.20$	0.002492	4.080052	120651	0.009967	0.009992	0.010190	0.010394
	$\epsilon_\infty = 2.0, \epsilon_1 = 0.40$	0.000617	0.237925	5443	0.002467	0.002492	0.002498	0.002610
	$\epsilon_\infty = 4.0, \epsilon_1 = 0.80$	0.000148	0.004939	24	0.000593	0.000617	0.000595	0.000659
$0.1\epsilon_\infty$	$\epsilon_\infty = 0.5, \epsilon_1 = 0.05$	0.039992	154	4941829	0.159967	0.159992	0.161191	0.161608
	$\epsilon_\infty = 1.0, \epsilon_1 = 0.10$	0.009992	20	620584	0.039967	0.039992	0.040201	0.040424
	$\epsilon_\infty = 2.0, \epsilon_1 = 0.20$	0.002492	1.255550	29356	0.009967	0.009992	0.010000	0.010130
	$\epsilon_\infty = 4.0, \epsilon_1 = 0.40$	0.000617	0.030494	156	0.002467	0.002492	0.002469	0.002560

From Table 6.1, one can notice that L-GRR presents the smallest variance values

Table 6.2: Numerical values of Eq. 2.5 (i.e., $Var^*[\hat{f}(v_i)]$) for the non-longitudinal GRR, OUE, and SUE protocols with different ϵ_∞ privacy guarantees.

ϵ_∞	GRR($c_j = 2$)	GRR($c_j = 32$)	GRR($c_j = 2^{10}$)	OUE	SUE
$\epsilon_\infty = 0.5$	**0.000392**	0.007520	0.243240	**0.001567**	**0.001592**
$\epsilon_\infty = 1.0$	**0.000092**	0.001108	0.034707	**0.000368**	**0.000392**
$\epsilon_\infty = 2.0$	**0.000018**	0.000092	0.002522	**0.000072**	**0.000092**
$\epsilon_\infty = 4.0$	0.000002	0.000003	0.000037	0.000008	0.000018

for binary attributes (i.e., when $c_j = 2$). On the other hand, L-GRR is also the most sensitive to change in privacy parameters ϵ_∞ and ϵ_1 when c_j is large, which leads to much higher variance than when using a non-longitudinal GRR in Table 6.2. Similar to non-longitudinal GRR, this increase in the variance is due to the number of values c_j, which decreases the probability p of reporting the true value. With two rounds of sanitization, it further deteriorates the accuracy of the L-GRR protocol getting to extremely high values, e.g., see L-GRR($c_j = 2^{10}$). Interestingly, when $c_j = 2$ in Table 6.1, the variance of L-GRR with $\epsilon_1 = 0.5\epsilon_\infty$ is a lagged version of the variance values given by the non-longitudinal GRR in Table 6.2. This effect is also observed for both L-SUE (cf. SUE in Table 6.2) and L-OSUE (cf. OUE in Table 6.2) protocols, which use symmetric probabilities on RR_2 (i.e., $p_2 + q_2 = 1$). We highlighted these values in **bold font**. However, for L-GRR, this is not true for other values of c_j, whose further analysis is beyond the scope of this chapter.

On the other hand, L-UE protocols avoid having a variance that depends on c_j by encoding the value into the unary representation, which results in a constant variance no matter the size of the attribute. To complement the results of Table 6.1, Fig. 6.3 illustrates numerical values of the approximate variance for L-UE protocols with $\epsilon_1 = \{0.3\epsilon_\infty, 0.6\epsilon_\infty\}$. With the four options I-IV analyzed, on high privacy regimes, L-OSUE and L-SUE have similar performance while *always* favoring the proposed L-OSUE one. On lower privacy regimes, our proposed protocols L-SOUE and L-OSUE have similar performance, which outperform both L-OUE and L-SUE protocols. As shown in our experiments, the L-OUE protocol has the worst performance among the four options analyzed, with the exception of high values for ϵ_∞ (see the plot on the bottom of Fig. 6.3), when it has performance superior or similar to L-SUE. Indeed, for L-OUE, selecting $p_2 = 1/2$ for the second sanitization step is too strict, which results in higher variance value. **Therefore, by comparing the approximate variances, the best option for L-UE protocols, in terms of utility, is starting with OUE and then with SUE as we propose in this chapter, i.e., L-OSUE.**

6.3.5/ THE *ALLOMFREE* ALGORITHM

Let $A = \{A_1, A_2, ..., A_d\}$ be a set of d attributes with domain size $\mathbf{c} = [c_1, c_2, ..., c_d]$, $\mathbb{A} = \{$*L-GRR*, *L-OSUE*$\}$ be a set of optimal longitudinal LDP protocols, and ϵ_∞ and ϵ_1 be the

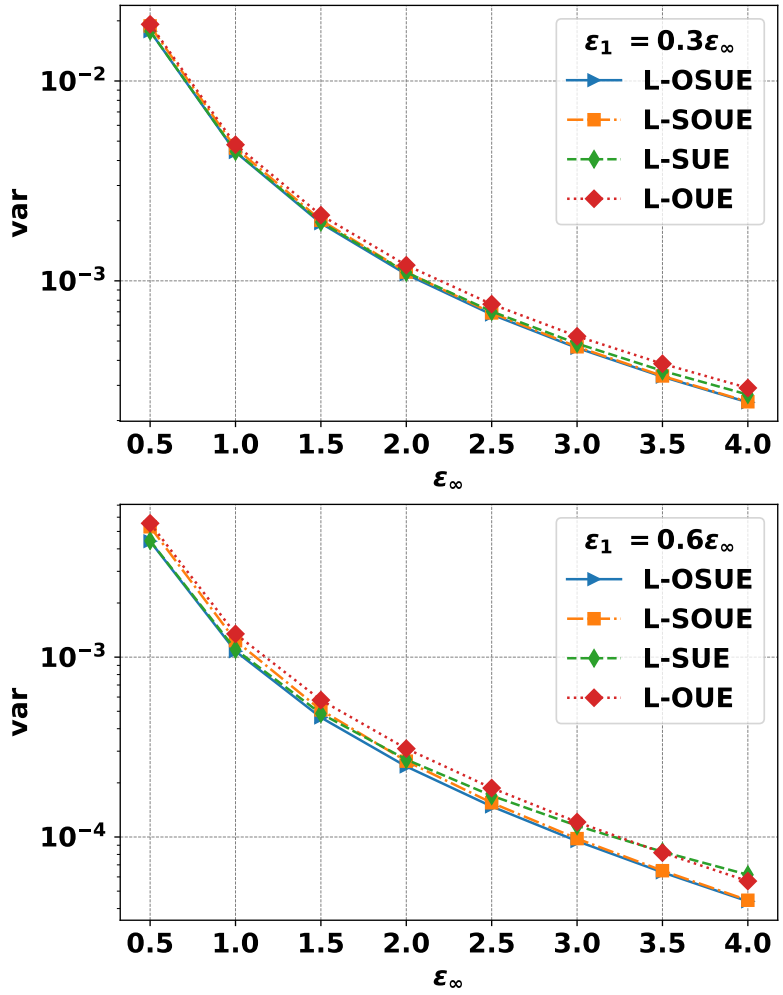

Figure 6.3: Numerical values of $Var^*[\hat{f}_L(v_i)]$ for L-UE protocols with $\epsilon_1 = 0.3\epsilon_\infty$ (plot on the top) and with $\epsilon_1 = 0.6\epsilon_\infty$ (plot on the bottom).

longitudinal and *single-report* privacy guarantees, respectively. Each user u_i, for $1 \leq i \leq n$, holds a tuple $\mathbf{v}^{(i)} = (v_1^{(i)}, v_2^{(i)}, ..., v_d^{(i)})$, i.e., a private value per attribute. From now on, we will simply omit the index notation $\mathbf{v}^{(i)}$ and use $\mathbf{v}$ in the analysis as we focus on one arbitrary user u_i here. For each attribute $j \in [1, d]$ (we slightly abuse the notation and use j for A_j) at time $t \in [1, \tau]$, the aggregator aims to estimate the frequencies of each value $v \in A_j$.

Client-Side. In a multidimensional setting with different domain sizes for each attribute, a dynamic selection of longitudinal LDP protocols is preferred. As mentioned in Section 6.2, we propose that each user randomly sample $r = Uniform(1, 2, ..., d)$ to select a single attribute A_r. Given c_r (the domain size), ϵ_∞, and ϵ_1, one calculates the parameters $fp_{L\text{-}GRR} = \{p_1, q_1, p_2, q_2\}$ and $fp_{L\text{-}OSUE} = \{p_1, q_1, p_2, q_2\}$, for L-GRR and L-OSUE, respectively (cf. Eq. (6.7) and Eq. (6.9)). Next, with $fp_{L\text{-}GRR}$ and $fp_{L\text{-}OSUE}$, one calculates the approximate variances $Var^*[\hat{f}_{L_{(L\text{-}GRR)}}]$ for L-GRR and $Var^*[\hat{f}_{L_{(L\text{-}OSUE)}}]$ for L-OSUE with Eq. (6.5). Lastly, to select L-GRR as the local randomizer, we are then left to evaluate if

$Var^*[\hat{f}_{L_{(L\text{-}GRR)}}] \leq Var^*[\hat{f}_{L_{(L\text{-}OSUE)}}]$. Therefore, the first round of sanitization ensures a *permanent memoization* B' that is always used for the second round of sanitization to generate B'' each time $t \in [1, \tau]$ the user will report the real value B. We call our solution Adaptive LDP for LOngitudinal and Multidimensional FREquency Estimates (ALLOMFREE), which is summarized in Algorithm 3 as a pseudocode.

Algorithm 3 User-side algorithm of ALLOMFREE.

Input : $\mathbf{v} = [v_1, v_2, ..., v_d]$, $\mathbf{c} = [c_1, c_2, ..., c_d]$, $\mathbb{A} = \{$*L-GRR*, *L-OSUE*$\}$, ϵ_∞, ϵ_1, number of reports τ.
$r \leftarrow Uniform(\{1, 2, ..., d\})$ ▷ Select attribute only once
$B \leftarrow v_r$
$fp_{L-GRR} \leftarrow p_1 = \frac{e^{\epsilon_\infty}}{e^{\epsilon_\infty}+k_r-1}, q_1 = \frac{1-p_1}{k_r-1}, p_2 = \frac{e^{\epsilon_1+\epsilon_\infty}-1}{-k_r e^{\epsilon_1}+(k_r-1)e^{\epsilon_\infty}+e^{\epsilon_1}+e^{\epsilon_1+\epsilon_\infty}-1}, q_2 = \frac{1-p_2}{k_r-1}$ ▷ Get p_2 and q_2 with Eq. (6.7).
$fp_{L-OSUE} \leftarrow p_1 = \frac{1}{2}, q_1 = \frac{1}{e^{\epsilon_\infty}+1}, p_2 = \frac{1-e^{\epsilon_1+\epsilon_\infty}}{e^{\epsilon_1}-e^{\epsilon_\infty}-e^{\epsilon_1+\epsilon_\infty}+1}, q_2 = 1 - p_2$ ▷ Get p_2 and q_2 with Eq. (6.9).
if $Var^*[\hat{f}_{L_{(L\text{-}GRR)}}](fp_{L-GRR}) \leq Var^*[\hat{f}_{L_{(L\text{-}OSUE)}}](fp_{L-OSUE})$: ▷ Check variances with Eq. (6.5))
  $\mathcal{A} \leftarrow$ L-GRR ▷ Select L-GRR as local randomizer
else
  $\mathcal{A} \leftarrow$ L-OSUE ▷ Select L-OSUE as local randomizer
$B' \leftarrow \mathcal{A}(B, \epsilon_\infty, c_r)$ ▷ First round of sanitization (permanent memoization)
for $t \in [1, \tau]$ **do**
  $B'' = \mathcal{A}(B', \epsilon_1, c_r)$ ▷ Second round of sanitization
end for
send : $(t, \langle r, B'' \rangle)$ for $t \in [1, \tau]$

The intuition of ALLOMFREE is as follows. By requiring each user to submit only 1 attribute with the whole privacy budget, it reduces both the variance incurred as well as the communication cost. Also, since we developed the calculus of the approximate variance in Eq. (6.5) for the proposed longitudinal protocols (L-GRR and L-OSUE), ALLOMFREE can adaptively select the protocol with a smaller variance value to optimize the data utility. Therefore, ALLOMFREE utilizes optimal solutions for both multidimensional and longitudinal data collection settings developed in Sections 6.2 and 6.3 of this manuscript, respectively.

Server-Side. On the server-side, for each attribute $j \in [1, d]$ at time $t \in [1, \tau]$, the estimated frequency $\hat{f}_L(v_i)$ that a value v_i occurs for $i \in [1, c_j]$ is calculated using Eq. (6.3).

Privacy analysis. On the one hand, according to the analysis in Subsections 6.3.2 and 6.3.3, Alg. 3 satisfies ϵ-LDP with upper ϵ_∞ (infinity reports) and lower ϵ_1 (a single report) bounds as it uses either L-GRR or L-OSUE to sanitize a single attribute per user. **Notice that, to ensure users' privacy over time and to avoid the sequential composition theorem [59], each user must always report the same unique attribute** A_r. In addition, the privacy of a user decreases gracefully according to the number of LDP reports $t \leq \tau$ that an adversary has gained access to, which is calculated as [195, 184]:

$$\epsilon_t = \ln\left(\frac{e^{\epsilon_\infty + t\epsilon_1} + 1}{e^{\epsilon_\infty} + e^{t\epsilon_1}}\right) \leq \min\{\epsilon_\infty, t\epsilon_1\}. \tag{6.10}$$

6.4/ Results and Discussion

In this section, we present the setup of our experiments in Section 6.4.1, the results with real-world data in Section 6.4.2, and a general discussion in Section 6.4.3 with related work and limitations.

6.4.1/ Setup of experiments

The main goal of our experiments is to evaluate the proposed longitudinal LDP protocols on multidimensional frequency estimates a single time, i.e., satisfying ϵ_1-LDP (as in [61, 200, 129], for example).

Environment. All algorithms were implemented in Python 3.8.8 with NumPy 1.19.5 and Numba 0.53.1 libraries. The codes we developed and used for all experiments are available in a Github repository[1]. In all experiments, we report average results over 100 runs as LDP algorithms are randomized.

Methods evaluated. We consider for evaluation the following solutions and protocols:

- Solution *Smp* (cf. Section 6.2), which randomly samples a single attribute to send with the whole privacy budget. We will experiment with the state-of-the-art protocols, namely, L-SUE and L-OUE, and with our extended protocols L-OSUE and L-SOUE;
- Our ALLOMFREE solution (cf. Alg. 3), which also randomly samples a single attribute to send with the whole privacy budget but adaptively select the optimal protocol, i.e., either L-GRR or L-OSUE.

Experimental evaluation and metrics. We vary the longitudinal privacy parameter in the range $\epsilon_\infty = [0.5, 1, ..., 3.5, 4]$ with $\epsilon_1 = \{0.3\epsilon_\infty, 0.6\epsilon_\infty\}$ to compare our experimental results with numerical ones from Section 6.3.4. Notice that this range of privacy guarantees is commonly used in the literature for multidimensional data (e.g., in [166] the range is $\epsilon = [0.5, ..., 4]$ and in [239] the range is $\epsilon = [0.1, ..., 10]$).

Since the estimator in Eq. (6.3) is unbiased (cf. Theorem 2), the variance of our protocols is equal to the MSE that is commonly used in practice as an accuracy metric [202, 203, 239, 224] (cf. Eq. (2.6)). So, to evaluate our results, we use the MSE metric averaged per

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

the number of attributes d **in a single data collection** $\tau = 1$**, i.e., with** ϵ_1**-LDP**. Thus, for each attribute j, we compute for each value $v_i \in A_j$ the estimated frequency $\hat{f}(v_i)$ and the real one $f(v_i)$ and calculate their differences. More precisely,

$$MSE_{avg} = \frac{1}{\tau} \sum_{t \in [1,\tau]} \frac{1}{d} \sum_{j \in [1,d]} \frac{1}{|A_j|} \sum_{v_i \in A_j} (f(v_i) - \hat{f}(v_i))^2. \tag{6.11}$$

Datasets. For ease of reproducibility, we conduct our experiments on four multidimensional open datasets. We briefly recall here the datasets from Section 3.3.6 and the generated one in Chapter 4.

- *Nursery.* A dataset from the UCI machine learning repository [96] with $d = 9$ categorical attributes and $n = 12960$ samples. The domain size of each attribute is $\mathbf{c} = [3, 5, 4, 4, 3, 2, 3, 3, 5]$, respectively.
- *Adult.* A dataset from the UCI machine learning repository [96] with $d = 9$ categorical attributes and $n = 45222$ samples after cleaning the data. The domain size of each attribute is $\mathbf{c} = [7, 16, 7, 14, 6, 5, 2, 41, 2]$, respectively.
- *MS-FIMU.* The dataset developed in Chapter 4 in which we select $d = 6$ categorical attributes (all static attributes, i.e., the dynamic 'Visit duration' attribute was not used). The domain size of each attribute is $\mathbf{c} = [3, 3, 8, 12, 37, 11]$ (cf. Section 4.4.2), respectively, and there are $n = 88935$ samples.
- *Census-Income.* A dataset from the UCI machine learning repository [96] with $d = 33$ categorical attributes and $n = 299285$ samples. The domain size of each attribute is $\mathbf{c} = [9, 52, 47, 17, 3, 7, 24, ..., 43, 5, 3, 3, 3, 2]$, respectively.

6.4.2/ RESULTS

Our experiments were conducted on four real-world datasets with varied parameters for n, d, and $\mathbf{c}$, which allowed evaluating our solutions more practically. Fig. 6.4 (*Nursery*), Fig. 6.5 (*Adult*), Fig. 6.6 (*MS-FIMU*), and Fig. 6.7 (*Census-Income*) illustrate for all evaluated protocols, averaged MSE_{avg} (y-axis) according to the longitudinal privacy parameter ϵ_∞ (x-axis) with $\epsilon_1 = 0.3\epsilon_\infty$ (right-side plot) and with $\epsilon_1 = 0.6\epsilon_\infty$ (left-side plot), respectively.

As one can notice in the results, for all datasets, ALLOMFREE consistently and considerably outperforms the state-of-the-art protocols, namely, L-SUE (a.k.a. Basic-RAPPOR) [61] and L-OUE (that uses OUE [108] twice). Indeed, the difference on performance between ALLOMFREE and the other longitudinal LDP protocols increases according to the privacy guarantees, i.e., for high ϵ_∞ and ϵ_1 values the gap is bigger. This is, first, because in all datasets there are attribute(s) with small domain size (e.g., $c_j = 2$ or

$c_j = 3$), in which L-GRR can provide smaller variance values than L-UE protocols (cf. Section 6.3.4). Secondly, by selecting adequately the probabilities p_1, q_1, p_2, q_2 for the L-UE protocol (i.e., L-OSUE) also optimizes data utility. Thus, since there is a way to measure the approximate variance of the extended protocols (i.e., Eq. (6.5)), given the sampled attribute, ALLOMFREE adaptively selects one of the optimized protocol (i.e., L-GRR or L-OSUE) whose smaller variance improves the data utility.

In addition, among the L-UE protocols applied individually, the experimental results with multidimensional data approximate the numerical results with a single attribute from Section 6.3.4. For instance, the proposed L-OSUE provides similar or improved performance than L-SUE while always outperforming L-OUE. Besides, L-SOUE always outperforms L-OUE too, achieving similar performance than L-OSUE and L-SUE in low privacy regimes (i.e., high ϵ values). As we have already shown in Section 6.3.4, even though OUE has higher utility than SUE for one-time collection [108], applying OUE twice does not provide higher utility.

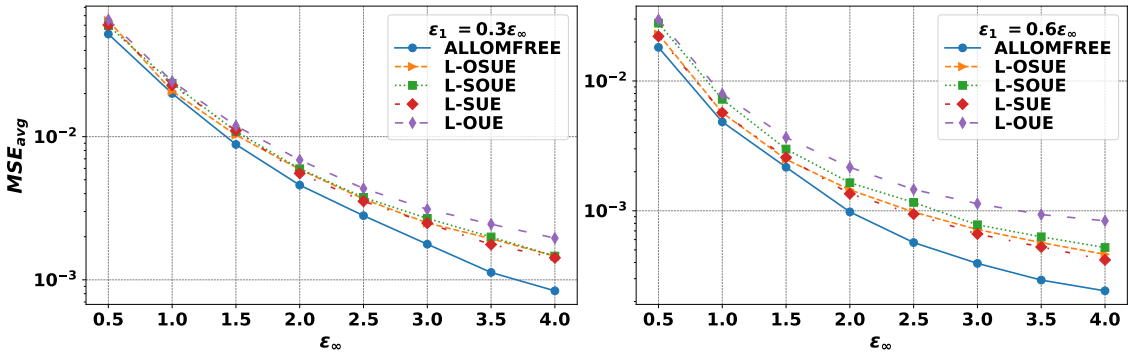

Figure 6.4: Averaged MSE varying ϵ_∞ with $\epsilon_1 = 0.3\epsilon_\infty$ (left-side plot) and with $\epsilon_1 = 0.6\epsilon_\infty$ (right-side plot) on the *Nursery* dataset.

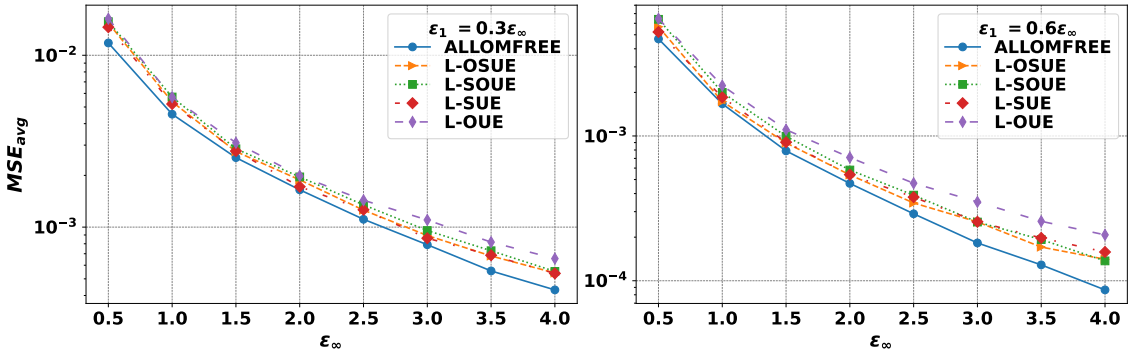

Figure 6.5: Averaged MSE varying ϵ_∞ with $\epsilon_1 = 0.3\epsilon_\infty$ (left-side plot) and with $\epsilon_1 = 0.6\epsilon_\infty$ (right-side plot) on the *Adult* dataset.

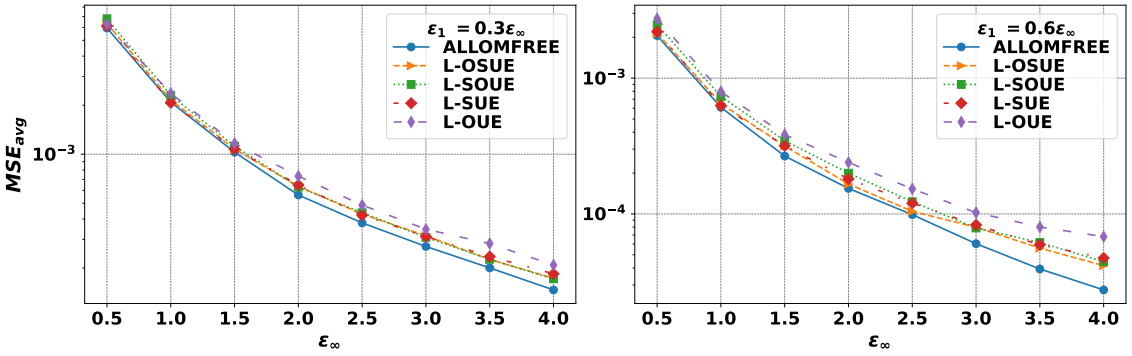

Figure 6.6: Averaged MSE varying ϵ_∞ with $\epsilon_1 = 0.3\epsilon_\infty$ (left-side plot) and with $\epsilon_1 = 0.6\epsilon_\infty$ (right-side plot) on the *MS-FIMU* dataset.

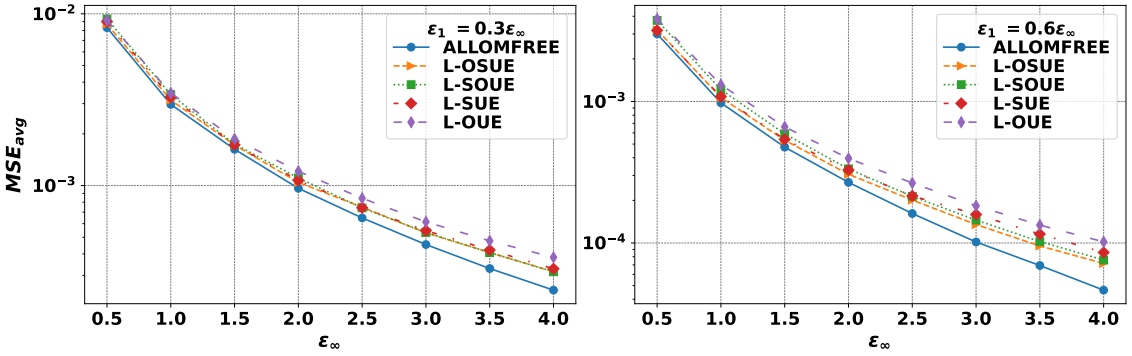

Figure 6.7: Averaged MSE varying ϵ_∞ with $\epsilon_1 = 0.3\epsilon_\infty$ (left-side plot) and with $\epsilon_1 = 0.6\epsilon_\infty$ (right-side plot) on the *Census-Income* dataset.

To complement the results of Figs. 6.4 – 6.7, Table 6.3 ($\epsilon_1 = 0.3\epsilon_\infty$) and Table 6.4 ($\epsilon_1 = 0.6\epsilon_\infty$) exhibit for all datasets and ϵ_∞ guarantees the following utility metrics:

$$\begin{aligned} \mathscr{U}_{\text{L-SUE}} &= \frac{MSE_{avg_{(\text{L-SUE})}} - MSE_{avg_{(\text{ALLOMFREE})}}}{MSE_{avg_{(\text{L-SUE})}}}, \\ \mathscr{U}_{\text{L-OUE}} &= \frac{MSE_{avg_{(\text{L-OUE})}} - MSE_{avg_{(\text{ALLOMFREE})}}}{MSE_{avg_{(\text{L-OUE})}}}, \end{aligned} \tag{6.12}$$

in which $\mathscr{U}_{\text{L-SUE}}$ and $\mathscr{U}_{\text{L-OUE}}$ represent the accuracy gain of ALLOMFREE over the state-of-the-art L-SUE and L-OUE protocols, respectively.

From Tables 6.3 and 6.4, one can notice that ALLOMFREE considerably improves the quality of the frequency estimates in comparison with the state-of-the-art L-SUE and L-OUE protocols. On average, ALLOMFREE improves the results of L-SUE at least 10% with the *MS-FIMU* dataset in Table 6.3 and at most 30.38% with the *Nursery* dataset in Table 6.4 for the privacy guarantees ϵ_∞ and ϵ_1 analyzed. Similarly, on average, ALLOMFREE improves the results of L-OUE at least 19.32% with the *MS-FIMU* dataset in Table 6.3 and at most 54.96% with the *Nursery* dataset in Table 6.4. The highest gain of

Table 6.3: Accuracy gain of ALLOMFREE over the state-of-the-art L-SUE and L-OUE protocols for all datasets with $\epsilon_1 = 0.3\epsilon_\infty$, measured with the $\mathscr{U}_{\text{L-SUE}}$ and $\mathscr{U}_{\text{L-OUE}}$ metrics expressed in %.

ϵ_∞	*Nursery*		*Adult*		*MS-FIMU*		*Census-Income*	
	$\mathscr{U}_{\text{L-SUE}}$	$\mathscr{U}_{\text{L-OUE}}$	$\mathscr{U}_{\text{L-SUE}}$	$\mathscr{U}_{\text{L-OUE}}$	$\mathscr{U}_{\text{L-SUE}}$	$\mathscr{U}_{\text{L-OUE}}$	$\mathscr{U}_{\text{L-SUE}}$	$\mathscr{U}_{\text{L-OUE}}$
0.5	13.51	20.63	19.03	27.73	3.03	5.43	7.84	9.48
1.0	12.36	17.75	12.77	20.44	1.01	11.57	9.21	14.08
1.5	19.95	25.86	8.47	18.01	4.13	11.55	5.82	12.92
2.0	17.18	33.24	4.11	17.16	13.22	23.44	10.06	20.41
2.5	20.70	35.40	11.93	22.54	10.41	22.25	12.77	23.15
3.0	28.69	42.98	8.35	28.22	13.07	21.56	17.07	26.21
3.5	36.19	54.02	18.97	32.02	14.78	29.10	22.02	30.96
4.0	41.24	57.16	19.81	34.25	20.38	29.64	24.99	35.60
Mean	23.73	35.88	12.93	25.05	10.00	19.32	13.72	21.60

Table 6.4: Accuracy gain of ALLOMFREE over the state-of-the-art L-SUE and L-OUE protocols for all datasets with $\epsilon_1 = 0.6\epsilon_\infty$, measured with the $\mathscr{U}_{\text{L-SUE}}$ and $\mathscr{U}_{\text{L-OUE}}$ metrics expressed in %.

ϵ_∞	*Nursery*		*Adult*		*MS-FIMU*		*Census-Income*	
	$\mathscr{U}_{\text{L-SUE}}$	$\mathscr{U}_{\text{L-OUE}}$	$\mathscr{U}_{\text{L-SUE}}$	$\mathscr{U}_{\text{L-OUE}}$	$\mathscr{U}_{\text{L-SUE}}$	$\mathscr{U}_{\text{L-OUE}}$	$\mathscr{U}_{\text{L-SUE}}$	$\mathscr{U}_{\text{L-OUE}}$
0.5	17.82	38.84	10.42	27.46	6.41	24.79	5.65	21.61
1.0	14.99	38.97	9.83	25.14	2.97	23.32	9.79	25.46
1.5	15.88	41.05	12.90	28.59	16.00	30.52	11.88	28.05
2.0	27.52	54.69	12.95	33.78	14.81	35.65	18.45	32.31
2.5	39.59	60.96	23.28	38.50	17.71	35.34	24.89	39.11
3.0	40.64	65.32	28.59	47.95	27.26	40.97	36.12	44.48
3.5	44.39	68.73	34.85	50.00	33.69	50.94	40.01	48.18
4.0	42.24	71.13	45.26	58.33	41.83	59.47	45.85	54.44
Mean	30.38	54.96	22.26	38.72	20.08	37.62	24.08	36.70

accuracy was about $\sim 71\%$, achieved with the *Nursery* dataset when $\epsilon_\infty = 4$ in Table 6.4 in comparison with the L-OUE protocol. Finally, as one can note, with higher values of ϵ_1, ALLOMFREE will provide much higher utility than the other protocols.

6.4.3/ Discussion and Related Work

Frequency estimation is a fundamental primitive in LDP and has received considerable attention for a single attribute in both theoretical and application perspectives [209, 204, 206] (see, e.g., [159, 139, 239, 150, 217, 108, 80, 116, 61, 95, 129, 172, 200, 117, 81, 168, 192, 107]). However, most studies for collecting multidimensional data with LDP mainly focused on numerical data [206] (e.g., cf. [83, 166, 123, 239]) or other complex tasks with categorical data, e.g., marginal estimation [233, 161, 138, 134, 78] and analytical/range queries [208, 207, 153, 147]. For instance, in [83, 166], the authors propose sampling-based LDP mechanisms for real-valued data (named Harmony and Piecewise

Mechanism) and applied these protocols in a multidimensional setting using state-of-the-art LDP mechanisms from [67, 108] for categorical data. Regarding multidimensional frequency estimates, in [108], the authors prove for the optimal local hashing protocol that sending 1 attribute with the whole privacy budget ϵ results in less variance than splitting the privacy budget for $d = 2$ attributes, i.e., with $\epsilon/2$. More generically, this is true for any number of attributes d for the GRR protocol, as we have shown analytically and empirically in Chapter 5, and for both OUE and SUE protocols, as shown in Section 6.2.

Besides, most frequency estimation academic literature focuses on single data collection. To address longitudinal data collections, in [61, 95, 184], the authors proposed LDP protocols based on two rounds of sanitization, i.e., *memoization*, which was also adopted in this chapter. In the literature, some works [129, 200] applied L-SUE (a.k.a. Basic-RAPPOR [61]) and L-OUE (i.e., OUE [108] two times) for longitudinal frequency estimates. However, rather than strictly using only SUE or OUE twice, we prove that the optimal combination is starting with OUE and then with SUE (i.e., L-OSUE). The privacy guarantees of chaining two LDP protocols has been further studied in [195, 184], which results in Eq. (6.10). Indeed, both "multiple" settings combined (i.e., many attributes and several collections throughout time), imposes several challenges, in which this paper, proposes the first solution named ALLOMFREE under LDP.

Indeed, both "multiple" settings combined (i.e., many attributes and several collections throughout time), imposes several challenges, in which this chapter, proposes the first solution named ALLOMFREE under ϵ-LDP. Yet, concerning the privacy guarantees of ALLOMFREE, the memoization step is certainly effective for longitudinal privacy to the cases where the true client's data does not vary (static) or vary very slowly or in an uncorrelated manner [61]. In many application scenarios, gender, age-ranges, nationality, and other demographic data are generally static or vary hardly ever. On the other hand, for dynamic attributes such as location or the time spent in the application, this is not the case. Therefore, for each different value, a new memoized value would be generated, thus accumulating the privacy budget ϵ_∞ by the sequential composition theorem [59].

6.5/ CONCLUSION

This chapter investigates the problem of collecting multidimensional data throughout time (i.e., longitudinal studies) for the fundamental task of frequency estimation under ϵ-LDP guarantees. We extended the analysis of three state-of-the-art LDP protocols, namely, GRR [80], OUE [108], and SUE [61] (cf. Section 2.4) for both longitudinal and multidimensional frequency estimates. On the one hand, for all three protocols, we theoretically prove that randomly sampling a single attribute per user improves data utility, which is an extension of common results in the LDP literature [184, 108, 208, 238, 138, 215].

On the other hand, in the literature, both SUE and OUE protocols have been extended (and also applied [129, 200]) to longitudinal studies based on the concept of *memoization* [61, 95], i.e., L-SUE and L-OUE, respectively. However, we numerically and experimentally show that combining both protocols provides higher data utility, i.e., starting with OUE and then with SUE (L-OSUE) minimizes the variance incurred rather than using SUE or OUE twice. In addition, for the first time, we also extended GRR for longitudinal studies (i.e., L-GRR), which provides higher data utility than the other protocols based on unary encoding for attributes with small domain sizes.

We also notice that in a multidimensional setting with different domain sizes for each attribute, a dynamic selection of longitudinal LDP protocols is preferred. Therefore, we also proposed a new solution named Adaptive LDP for LOngitudinal and Multidimensional FREquency Estimates (ALLOMFREE), which combines all the aforementioned results. More specifically, ALLOMFREE randomly samples a single attribute to send with the whole privacy budget and adaptively selects the optimal protocol, i.e., either L-GRR or L-OSUE.

To validate our proposal, we conducted a comprehensive and extensive set of experiments on four real-world open datasets. Under the same privacy guarantee, results show that ALLOMFREE consistently and considerably outperforms the state-of-the-art L-SUE [61] and L-OUE [108] protocols in the quality of the frequency estimates, with a gain of accuracy, on average, ranging from 10% up to 55%.

Lastly, we highlight that ALLOMFREE is based on the multidimensional *Smp* solution, which randomly samples a single attribute out of d ones to send it with ϵ-LDP. However, aggregators (who are also seen as attackers) are aware of the sampled attribute and its LDP value, which is protected by a "less strict" e^{ϵ} probability bound (rather than $e^{\epsilon/d}$). Indeed, in some cases, using the *Smp* solution may be "unfair" with some users, e.g., users that randomly sample a demographic attribute (e.g., age) might be less concerned to report their data than those whose sampled attribute is socially "more" sensitive (e.g., disease, location, most common web page). Investigating how to deal with this "unfair" issue on multidimensional frequency estimates is the main *goal* of the next Chapter 7.

7

Multidimensional Frequency Estimates With LDP: Privacy Focus

In Chapter 6, we tackled both multidimensional and longitudinal settings for the fundamental task of frequency estimation under ϵ-LDP guarantees. In this chapter, we continue contributing to the **theoretical** aspect dedicating our efforts to the multidimensional setting only. Indeed, the sampling-based solution for multidimensional frequency estimates used in Chapters 5 and 6 (i.e., *Smp*), focuses on optimizing **the utility**. However, this solution considers that all attributes have equal weight in terms of privacy, which (generally) is not the case in real life. For example, in health data collection, people who randomly sample the disease attribute might hesitate to share their data in comparison with others that randomly sample, e.g., age. This idea extends to other application scenarios, e.g., in software monitoring applications with the "favorite webpage" attribute, and so on. Therefore, in this chapter, we propose a solution for multidimensional frequency estimates under ϵ-LDP guarantees, which improves the **privacy** of users while providing **the same or better performance** than the state-of-the-art *Smp* solution.

7.1/ Introduction

We start recalling the problem statement here. As in previous chapters, we assume there are d attributes $A = \{A_1, A_2, ..., A_d\}$, where each attribute A_j with a discrete domain has a specific number of values $|A_j| = c_j$. Each user u_i for $i \in \{1, 2, ..., n\}$ has a tuple $\mathbf{v}^{(i)} = (v_1^{(i)}, v_2^{(i)}, ..., v_d^{(i)})$, where $v_j^{(i)}$ represents the value of attribute A_j in record $\mathbf{v}^{(i)}$. Thus, for each attribute A_j, the analyzer's goal is to estimate a c_j-bins histogram, including the frequency of all values in A_j.

As presented in Chapters 5 and 6, there are mainly two solutions for satisfying LDP by

randomizing $\mathbf{v}$, namely, *Spl* and *Smp*. We will also omit the index notation $\mathbf{v}^{(i)}$ and use $\mathbf{v}$ in the analysis as we focus on one arbitrary user u_i here. Although the *Smp* solution adds sampling error, in the literature [166, 83, 108, 215, 239, 208, 238, 184, 88] and in previous Chapters 5 and 6, *Smp* has proven to provide higher data utility than the former *Spl* solution.

However, as aforementioned, aggregators (who are also seen as attackers) are aware of the sampled attribute and its LDP value, which is protected by a "less strict" e^{ϵ} probability bound (rather than $e^{\epsilon/d}$). In other words, while both solutions provide ϵ-LDP, we argue that using the *Smp* solution may be unfair with some users. For instance, on collecting multidimensional health records (i.e., demographic and clinical data), users that randomly sample a demographic attribute (e.g., gender) might be less concerned to report their data than those whose sampled attribute is "disease" (e.g., if positive for human immunodeficiency viruses - HIV).

This way, there is a privacy-utility trade-off between the *Spl* and *Smp* solutions. With these elements in mind, we formulate the **problematic of this chapter** as: *For the same privacy budget ϵ, is there a solution for multidimensional frequency estimates that provides better data utility than Spl and more protection than Smp?*

Thus, we intend to solve the aforementioned problematic by answering the following question: *What if the sampling result (i.e., the selected attribute) was **not** disclosed with the aggregator?* Thus, since the sampling step randomly selects an attribute $j \in [1, d]$ (we slightly abuse the notation and use j for A_j), we propose that users **add uncertainty about the sampled attribute** through generating $d-1$ *fake data*, i.e., one for each non-sampled attribute.

We call our solution *Random Sampling plus Fake Data (RS+FD)*. On the one hand, since RS+FD introduces some *uncertainty* in the view of the aggregator, we remarked that users' *privacy* is amplified by sampling [25, 43, 118, 173, 32]. Besides, we integrate two state-of-the-art LDP protocols, namely, GRR [80] and OUE [108] for single attribute frequency estimation, both presented in Section 2.4, into our RS+FD solution to propose four protocols. We demonstrate through experimental validations using four real-world datasets the advantages of our protocols with RS+FD over the state-of-the-art *Spl* and *Smp* solutions.

The rest of this chapter is organized as follows. In Section 7.2, we introduce our RS+FD solution, the integration of state-of-the-art LDP mechanisms within RS+FD, and their analysis. In Section 7.3, we present experimental results. Lastly, in Section 7.5, we present the concluding remarks. The proposed RS+FD solution in Section 7.2 and the results presented in Section 7.3 were published in a full paper [213] at the 30th International Conference on Information and Knowledge Management (CIKM 2021).

7.2/ RANDOM SAMPLING PLUS FAKE DATA (RS+FD)

In this section, we present the overview of our RS+FD solution (Section 7.2.1), and the integration of the local randomizers presented in Section 2.4 within RS+FD (Subsections 7.2.2, 7.2.3, and 7.2.4).

7.2.1/ OVERVIEW OF RS+FD

Fig. 7.1 illustrates the overview of our proposed RS+FD solution in comparison with the aforementioned known solutions, namely, *Spl* and *Smp*, which is detailed in the following.

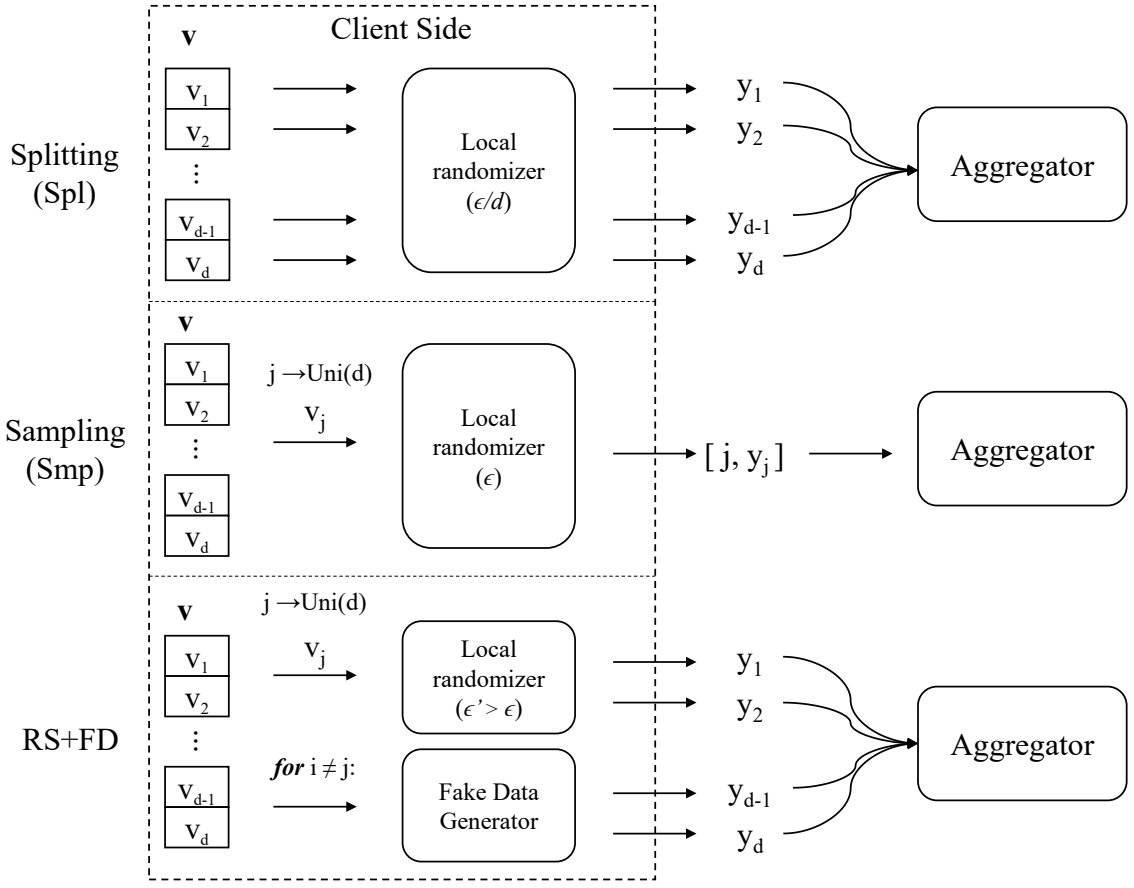

Figure 7.1: Overview of our random sampling plus fake data (RS+FD) solution in comparison with two known solutions, namely, *Spl* and *Smp*, where $Uni(d) = Uniform(\{1, 2, ..., d\})$.

We consider the local DP model, in which there are two entities, namely, users and the aggregator (an untrusted curator). Let n be the total number of users, d be the total number of attributes, $\mathbf{c} = [c_1, c_2, ..., c_d]$ be the domain size of each attribute, $\mathcal{A}$ be a local randomizer, and ϵ be the whole privacy budget. Each user holds a tuple $\mathbf{v} = (v_1, v_2, ..., v_d)$, i.e., a private value per attribute.

Client-Side. The client-side is split into two steps, namely, local randomization and fake data generation (cf. Fig. 7.1). Initially, each user samples a unique attribute j uniformly at random and applies an LDP mechanism to its value v_j. Indeed, RS+FD is generic to be applied with any existing LDP mechanisms (e.g., GRR [80], UE- or hash-based protocols [61, 108], Hadamard Response [139]). Next, for each $d-1$ non-sampled attribute i, the user generates one random fake data. Finally, each user sends the (LDP or fake) value of each attribute to the aggregator, i.e., a tuple $\mathbf{y} = (y_1, y_2, ..., y_d)$. This way, the sampling result is not disclosed with the aggregator and, thus, an amplified privacy budget $\epsilon' \geq \epsilon$ can be used. In summary, Alg. 4 exhibits the pseudocode of our RS+FD solution.

Algorithm 4 Random Sampling plus Fake Data (RS+FD)

Input : tuple $\mathbf{v} = (v_1, v_2, ..., v_d)$, domain size of attributes $\mathbf{c} = [c_1, c_2, ..., c_d]$, privacy parameter ϵ, local randomizer $\mathcal{A}$.
Output : sanitized tuple $\mathbf{y} = (y_1, y_2, ..., y_d)$.

$\epsilon' = \ln(d \cdot (e^{\epsilon} - 1) + 1)$ ▷ amplification by sampling [43]
$j \leftarrow Uniform(\{1, 2, ..., d\})$ ▷ Selection of attribute to sanitize
$B_j \leftarrow v_j$
$y_j \leftarrow \mathcal{A}(B_j, c_j, \epsilon')$ ▷ sanitize data of the sampled attribute
for $i \in \{1, 2, ..., d\} \setminus \{j\}$ **do** ▷ non-sampled attributes
  $y_i \leftarrow Uniform(\{1, ..., c_i\})$ ▷ generate fake data
end for
return : $\mathbf{y} = (y_1, y_2, ..., y_d)$ ▷ sampling result is not disclosed

Aggregator. For each attribute $j \in [1, d]$, the aggregator performs frequency (or histogram) estimation on the collected data by removing bias introduced by the local randomizer and fake data.

Privacy analysis. Let $\mathcal{A}$ be any existing LDP mechanism, Algorithm 4 satisfies ϵ-LDP, in a way that $\epsilon' = \ln(d \cdot (e^{\epsilon} - 1) + 1)$. Indeed, we observe that our scenario is equivalent to sampling a dataset $\mathcal{D}$ without replacement with sampling rate $\beta = \frac{1}{d}$ in the centralized setting of DP, which enjoys privacy amplification (cf. Section 2.3.3). More specifically, let a trusted curator in the centralized DP setting randomly split a dataset $\mathcal{D}$ in d disjoint subsets $D_1, D_2, ..., D_d$, i.e., each with n/d non-overlapping users. Next, let the trusted curator perform frequency estimation in each subset $D \in \mathcal{D}$ with ϵ'-DP. Therefore, invoking Theorem 1 (amplification by sampling) and Proposition 3 (parallel composition), all frequency estimation queries satisfy ϵ-DP with $\epsilon = \ln\left(1 + \beta(e^{\epsilon'} - 1)\right)$ where $\beta = 1/d$. In our case, with the local model, users sanitize their data locally with a DP model. This way, to satisfy ϵ-LDP with RS+FD, an amplified privacy parameter $\epsilon' \geq \epsilon$ can be used.

Limitations. Similar to other sampling-based methods for collecting multidimensional data under LDP [123, 83, 166, 239], our RS+FD solution also entails *sampling error*, which is due to observing a sample instead of the entire population. In addition, in comparison with the *Smp* solution, RS+FD requires more computation on the user side be-

cause of the fake data generation part. Yet, communication cost is still equal to the *Spl* solution, i.e., each user sends one message per attribute. Lastly, while RS+FD utilizes an amplified $\epsilon' \geq \epsilon$, there is also bias generated from uniform fake data that may require a sufficient number of users n to eliminate the noise.

7.2.2/ RS+FD WITH GRR

Client side. Integrating GRR as the local randomizer $\mathcal{A}$ into Alg. 4 (RS+FD[GRR]) requires no modification. Initially, on the client-side, each user randomly samples an attribute j. Next, the value v_j is sanitized with GRR (cf. Section 2.4.2) using the size of the domain c_j and the privacy parameter ϵ'. In addition, for each non-sampled $d-1$ attribute i, the user also generates fake data uniformly at random according to the domain size c_i. Lastly, the user transmits the sanitized tuple $\mathbf{y}$, which includes the LDP value of the true data "hidden" among fake data. Visually, Fig. 7.2 illustrates the probability tree of the RS+FD[GRR] protocol.

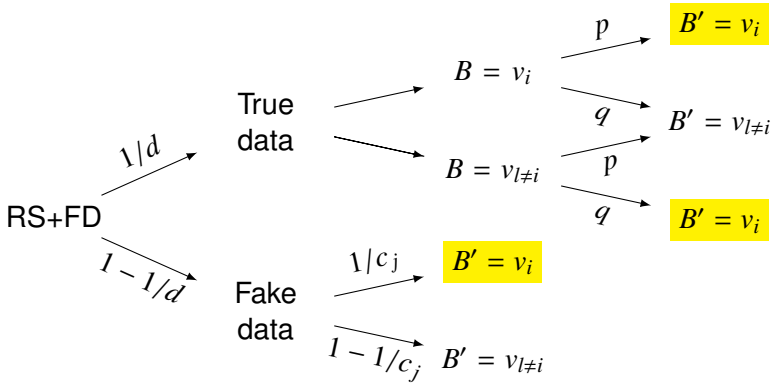

Figure 7.2: Probability tree for the RS+FD[GRR] protocol.

Aggregator RS+FD[GRR]. On the server-side, according to the probability tree in Fig. 7.2, for each attribute $j \in [1, d]$, the aggregator estimates $\hat{f}(v_i)$ for the frequency of each value $i \in [1, c_j]$ as:

$$\hat{f}(v_i) = \frac{N_i d c_j - n(d - 1 + q c_j)}{n c_j (p - q)}, \tag{7.1}$$

in which N_i is the number of times the value v_i has been reported, $p = \frac{e^{\epsilon'}}{e^{\epsilon'} + c_j - 1}$, and $q = \frac{1-p}{c_j - 1}$.

Theorem 4. *For $j \in [1, d]$, the estimation result $\hat{f}(v_i)$ in Eq. (7.1) is an unbiased estimation of $f(v_i)$ for any value $v_i \in A_j$.*

Proof.

$$E[\hat{f}(v_i)] = E\left[\frac{N_i d c_j - n(d-1+qc_j)}{nc_j(p-q)}\right]$$
$$= \frac{d}{n(p-q)}E[Ni] - \frac{d-1+qc_j}{c_j(p-q)}.$$

Let us focus on

$$E[N_i] = \frac{1}{d}\left(pnf(v_i) + q(n - nf(v_i))\right) + \frac{d-1}{dc_j}n$$
$$= \frac{n}{d}\left(f(v_i)(p-q) + q + \frac{d-1}{c_j}\right).$$

Thus,

$$E[\hat{f}(v_i)] = f(v_i).$$

□

Theorem 5. *The variance of the estimation in Eq.* (7.1) *is:*

$$\mathrm{VAR}(\hat{f}(v_i)) = \frac{d^2\gamma(1-\gamma)}{n(p-q)^2}, \text{ where}$$
$$\gamma = \frac{1}{d}\left(q + f(v_i)(p-q) + \frac{(d-1)}{c_j}\right). \tag{7.2}$$

Proof. Thanks to Eq. (7.1) we have

$$\mathrm{VAR}(\hat{f}(v_i)) = \frac{\mathrm{VAR}(N_i)d^2}{n^2(p-q)^2}.$$

Since N_i is the number of times value v_i is observed, it can be defined as $N_i = \sum_{z=1}^{n} X_z$ where X_z is equal to 1 if the user z, $1 \le z \le n$ reports value v_i, and 0 otherwise. We thus have $\mathrm{VAR}(N_i) = \sum_{z=1}^{n} \mathrm{VAR}(X_z) = n\,\mathrm{VAR}(X)$, since all the users are independent. According to the probability tree in Fig. 7.2,

$$P(X=1) = P(X^2=1) = \gamma = \frac{1}{d}\left(q + f(v_i)(p-q) + \frac{(d-1)}{c_j}\right).$$

We thus have $\mathrm{VAR}(X) = \gamma - \gamma^2 = \gamma(1-\gamma)$ and, finally,

$$\mathrm{VAR}(\hat{f}(v_i)) = \frac{d^2\gamma(1-\gamma)}{n(p-q)^2}.$$

□

7.2.3/ RS+FD WITH OUE

Client side. To use UE-based protocols (OUE in our work) as local randomizer $\mathcal{A}$ in Alg. 4, there is, first, a need to define the fake data generation procedure. We propose two solutions: (i) RS+FD[OUE-z] in Alg. 5, which applies OUE to $d-1$ zero-vectors (i.e., vectors with only zeros, e.g., $[0,0,...,0,0]$), and (ii) RS+FD[OUE-r] in Alg. 6, which applies OUE to $d-1$ one-hot-encoded fake data (uniform at random). Visually, Figs. 7.3 and 7.4 illustrate the probability trees of the RS+FD[OUE-z] and RS+FD[OUE-r] protocols, respectively.

Algorithm 5 RS+FD[OUE-z]

Input : tuple $\mathbf{v} = (v_1, v_2, ..., v_d)$, domain size of attributes $\mathbf{c} = [c_1, c_2, ..., c_d]$, privacy parameter ϵ, local randomizer OUE.

Output : sanitized tuple $\mathbf{B'} = (B'_1, B'_2, ..., B'_d)$.

$\epsilon' = \ln(d \cdot (e^{\epsilon} - 1) + 1)$ ▷ amplification by sampling [43]

$j \leftarrow Uniform(\{1, 2, ..., d\})$ ▷ Selection of attribute to sanitize

$B_j = Encode(v_j) = [0, 0, ..., 1, 0, ...0]$ ▷ one-hot-encoding

$B'_j \leftarrow OUE(B_j, \epsilon')$ ▷ sanitize real data with OUE

for $i \in \{1, 2, ..., d\} \setminus \{j\}$ **do** ▷ non-sampled attributes

  $B_i \leftarrow [0, 0, ..., 0]$ ▷ initialize zero-vectors

  $B'_i \leftarrow OUE(B_i, \epsilon')$ ▷ randomize zero-vector with OUE

end for

return : $\mathbf{B'} = (B'_1, B'_2, ..., B'_d)$ ▷ sampling result is not disclosed

Algorithm 6 RS+FD[OUE-r]

Input : tuple $\mathbf{v} = (v_1, v_2, ..., v_d)$, domain size of attributes $\mathbf{c} = [c_1, c_2, ..., c_d]$, privacy parameter ϵ, local randomizer OUE.

Output : sanitized tuple $\mathbf{B'} = (B'_1, B'_2, ..., B'_d)$.

$\epsilon' = \ln(d \cdot (e^{\epsilon} - 1) + 1)$ ▷ amplification by sampling [43]

$j \leftarrow Uniform(\{1, 2, ..., d\})$ ▷ Selection of attribute to sanitize

$B_j = Encode(v_j) = [0, 0, ..., 1, 0, ...0]$ ▷ one-hot-encoding

$B'_j \leftarrow OUE(B_j, \epsilon')$ ▷ sanitize real data with OUE

for $i \in \{1, 2, ..., d\} \setminus \{j\}$ **do** ▷ non-sampled attributes

  $y_i \leftarrow \mathit{Uniform}(\{1, ..., c_i\})$ ▷ generate fake data

  $B_i \leftarrow Encode(y_i)$ ▷ one-hot-encoding

  $B'_i \leftarrow OUE(B_i, \epsilon')$ ▷ randomize fake data with OUE

end for

return : $\mathbf{B'} = (B'_1, B'_2, ..., B'_d)$ ▷ sampling result is not disclosed

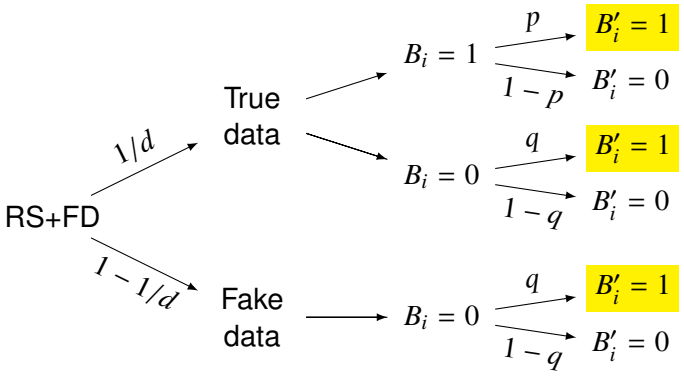

Figure 7.3: Probability tree for the RS+FD[OUE-z] protocol.

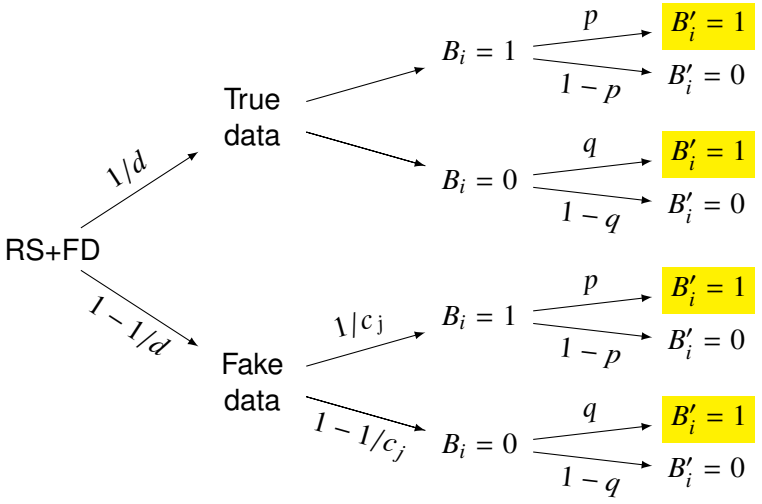

Figure 7.4: Probability tree for the RS+FD[OUE-r] protocol.

Aggregator RS+FD[OUE-z]. On the server-side, if fake data are generated with OUE applied to zero-vectors as in Alg. 5, according to the probability tree in Fig. 7.3, for each attribute $j \in [1, d]$, the aggregator estimates $\hat{f}(v_i)$ for the frequency of each value $i \in [1, c_j]$ as:

$$\hat{f}(v_i) = \frac{d(N_i - nq)}{n(p - q)}, \tag{7.3}$$

in which N_i is the number of times the value v_i has been reported, n is the total number of users, $p = \frac{1}{2}$, and $q = \frac{1}{e^{\epsilon'}+1}$.

Theorem 6. *For $j \in [1, d]$, the estimation result $\hat{f}(v_i)$ in Eq. (7.3) is an unbiased estimation of $f(v_i)$ for any value $v_i \in A_j$.*

Proof.

$$E[\hat{f}(v_i)] = E\left[\frac{d(N_i - nq)}{n(p-q)}\right] = \frac{d(E[N_i] - nq)}{n(p-q)}$$
$$= \frac{d}{n(p-q)}E[N_i] - \frac{dq}{p-q}.$$

We have successively

$$E[N_i] = \frac{n}{d}\left(pf(v_i) + q(1 - f(v_i))\right) + \frac{(d-1)nq}{d}$$
$$= \frac{n}{d}\left(f(v_i)(p-q) + dq\right).$$

Thus,

$$E[\hat{f}(v_i)] = f(v_i).$$

□

Theorem 7. *The variance of the estimation in Eq. (7.3) is:*

$$\begin{aligned} \mathrm{VAR}(\hat{f}(v_i)) &= \frac{d^2\gamma(1-\gamma)}{n(p-q)^2}, \text{ where} \\ \gamma &= \frac{1}{d}\left(dq + f(v_i)(p-q)\right). \end{aligned} \tag{7.4}$$

The proof for Theorem 7 follows the *Proof* of Theorem 5 and is omitted here. In this case, γ follows the probability tree in Fig. 7.3.

Aggregator RS+FD[OUE-r]. Otherwise, if fake data are generated with OUE applied to one-hot-encoded random data as in Alg. 6, according to the probability tree in Fig. 7.4, for each attribute $j \in [1, d]$, the aggregator estimates $\hat{f}(v_i)$ for the frequency of each value $i \in [1, c_j]$ as:

$$\hat{f}(v_i) = \frac{N_i d c_j - n\left[qc_j + (p-q)(d-1) + qc_j(d-1))\right]}{nc_j(p-q)}, \tag{7.5}$$

in which N_i is the number of times the value v_i has been reported, $p = \frac{1}{2}$, and $q = \frac{1}{e^{\epsilon'}+1}$.

Theorem 8. *For $j \in [1, d]$, the estimation result $\hat{f}(v_i)$ in Eq. (7.5) is an unbiased estimation of $f(v_i)$ for any value $v_i \in A_j$.*

Proof.

$$E[\hat{f}(v_i)] = E\left[\frac{N_i d c_j - n\left[qc_j + (p-q)(d-1) + qc_j(d-1))\right]}{nc_j(p-q)}\right]$$
$$= \frac{dE[N_i]}{n(p-q)} - \frac{(p-q)(d-1) + qdc_j}{c_j(p-q)}.$$

We have successively

$$\begin{aligned} E[N_i] &= \frac{n}{d}\left(pf(v_i) + q(1 - f(v_i))\right) + \frac{n(d-1)}{d}\left(\frac{p}{c_j} + \frac{c_j - 1}{c_j}q\right) \\ &= \frac{n}{d}\left(f(v_i)(p-q) + q\right)\right) + \frac{n(d-1)}{dc_j}(p - q + c_j q). \end{aligned}$$

Thus,

$$E[\hat{f}(v_i)] = f(v_i).$$

□

Theorem 9. *The variance of the estimation in Eq.* (7.5) *is:*

$$\begin{aligned} \mathrm{VAR}(\hat{f}(v_i)) &= \frac{d^2\gamma(1-\gamma)}{n(p-q)^2}, \textit{ where} \\ \gamma &= \frac{1}{d}\left(q + f(v_i)(p-q) + \frac{(d-1)}{c_j}\Big(p + (c_j - 1)q\Big)\right). \end{aligned} \tag{7.6}$$

The proof for Theorem 9 follows the *Proof* of Theorem 5 and is omitted here. In this case, γ follows the probability tree in Fig. 7.4.

7.2.4/ Analytical analysis: RS+FD with ADP

As shown in Chapter 6, in a multidimensional setting with different domain sizes for each attribute, a dynamic selection of LDP mechanisms is preferred. In this chapter, we also analyze the *approximate variances* VAR_1 for RS+FD[GRR] in Eq. (7.2) and VAR_2 for RS+FD[OUE-z] in Eq. (7.4), in which $f(v_i) = 0$. Assume there are $d \geq 2$ attributes with domain size $\mathbf{c} = [c_1, c_2, ..., c_d]$ and a privacy budget ϵ'. For each attribute j with domain size c_j, to select RS+FD[GRR], we are then left to evaluate if $\mathrm{VAR}_1 \leq \mathrm{VAR}_2$. This is equivalent to check whether,

$$\frac{d^2\gamma_1(1-\gamma_1)}{n(p_1-q_1)^2} - \frac{d^2\gamma_2(1-\gamma_2)}{n(p_2-q_2)^2} \leq 0, \tag{7.7}$$

in which $p_1 = \frac{e^{\epsilon'}}{e^{\epsilon'}+c_j-1}$, $q_1 = \frac{1-p_1}{c_j-1}$, $p_2 = \frac{1}{2}$, $q_2 = \frac{1}{e^{\epsilon'}+1}$, $\gamma_1 = \frac{1}{d}\left(q_1 + \frac{d-1}{c_j}\right)$, and $\gamma_2 = q_2$. **In other words, if Eq. (7.7) is verified, the utility loss is lower with RS+FD[GRR]; otherwise, RS+FD[OUE-z] should be selected. Throughout this chapter, we will refer to this dynamic selection of our protocols as RS+FD[ADP].**

For the sake of illustration, Fig. 7.5 illustrates a 3D visualization of $\frac{d^2\gamma_1(1-\gamma_1)}{n(p_1-q_1)^2} - \frac{d^2\gamma_2(1-\gamma_2)}{n(p_2-q_2)^2}$, i.e., the left side of Eq. (7.7), by fixing $\epsilon' = \ln(3)$ and $n = 10000$, and by varying $d \in [2, 10]$ and

$c_j \in [2, 20]$, which are common values for real-world datasets (cf. Section 7.3.1). In this case, one can notice in Fig. 7.5 that neither RS+FD[GRR] nor RS+FD[OUE-z] will always provide the lowest variance value, which reinforces the need for an adaptive mechanism. For instance, with the selected parameters, for lower values of c_j, RS+FD[GRR] can provide lower estimation errors even if d is large. On the other hand, as soon as the domain size starts to grow, e.g., $c_j \geq 10$, one is better off with RS+FD[OUE-z] even for small values of $d \geq 3$, as its variance in Eq. (7.4) does not depend on c_j.

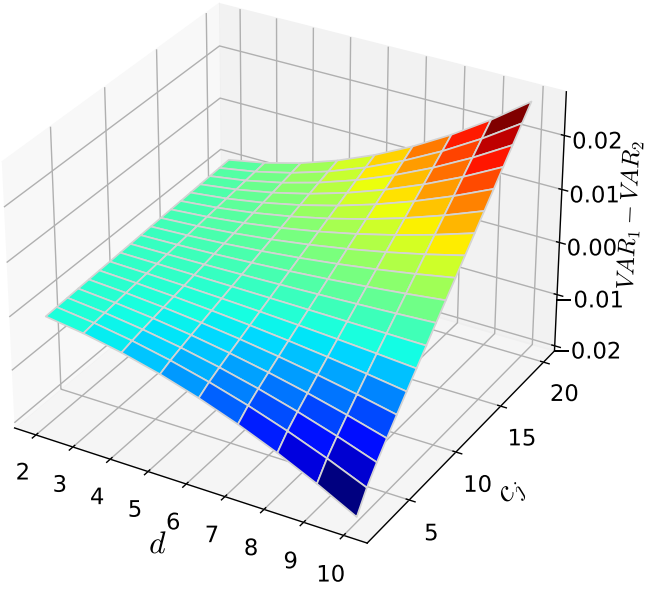

Figure 7.5: Analytical evaluation of Eq. (7.7) that allows a dynamic selection between RS+FD[GRR] with variance VAR_1 and RS+FD[OUE-z] with variance VAR_2. Parameters were set as $\epsilon' = \ln(3)$, $n = 10000$, $d \in [2, 10]$, and $c_j \in [2, 20]$.

7.3/ Experimental Validation

In this section, we present the setup of our experiments in Section 7.3.1, the results with synthetic data in Section 7.3.2, and the results with real-world data in Section 7.3.3.

7.3.1/ Setup of Experiments

Environment. All algorithms were implemented in Python 3.8.5 with NumPy 1.19.5 and Numba 0.53.1 libraries. The codes we developed and used for all experiments are available in a Github repository[1]. In all experiments, we report average results over 100 runs as LDP algorithms are randomized.

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

Synthetic datasets. Our first set of experiments are conducted on six synthetic datasets. The distribution of values in each attribute follows an uniform distribution, for all synthetic datasets.

- For the first two synthetic datasets, we fix the number of attributes $d = 5$ and the domain size of each attribute as $\mathbf{c} = [10, 10, ..., 10]$ (uniform), and vary the number of users as $n = 50000$ and $n = 500000$.
- Similarly, for the third and fourth synthetic datasets, we fix the number of attributes $d = 10$ and the domain size of each attribute as $\mathbf{c} = [10, 10, ..., 10]$ (uniform), and vary the number of users as $n = 50000$ and $n = 500000$.
- Lastly, for the fifth and sixth synthetic datasets, we fix the number of users as $n = 500000$. Next, we set the number of attributes $d = 10$ with domain size of each attribute as $\mathbf{c} = [10, 20, ..., 90, 100]$ for one dataset, and we set the number of attributes $d = 20$ with domain size of each attribute as $\mathbf{c} = [10, 10, 20, 20, ..., 100, 100]$ for the other.

Real-world datasets. In addition, we also conduct experiments on four real-world open datasets with non-uniform distributions. We briefly recall here the datasets from Section 3.3.6 and the generated one in Chapter 4.

- *Nursery.* A dataset from the UCI machine learning repository [96] with $d = 9$ categorical attributes and $n = 12960$ samples. The domain size of each attribute is $\mathbf{c} = [3, 5, 4, 4, 3, 2, 3, 3, 5]$, respectively.
- *Adult.* A dataset from the UCI machine learning repository [96] with $d = 9$ categorical attributes and $n = 45222$ samples after cleaning the data. The domain size of each attribute is $\mathbf{c} = [7, 16, 7, 14, 6, 5, 2, 41, 2]$, respectively.
- *MS-FIMU.* The dataset developed in Chapter 4 in which we select $d = 6$ categorical attributes (all static attributes, i.e., the dynamic 'Visit duration' attribute was not used). The domain size of each attribute is $\mathbf{c} = [3, 3, 8, 12, 37, 11]$ (cf. Section 4.4.2), respectively, and there are $n = 88935$ samples.
- *Census-Income.* A dataset from the UCI machine learning repository [96] with $d = 33$ categorical attributes and $n = 299285$ samples. The domain size of each attribute is $\mathbf{c} = [9, 52, 47, 17, 3, 7, 24, ..., 43, 5, 3, 3, 3, 2]$, respectively.

Evaluation and metrics. We vary the privacy parameter in a logarithmic range as $\epsilon = [\ln(2), \ln(3), ..., \ln(7)]$, which is within range of values experimented in the literature for multidimensional data (e.g., in [166] the range is $\epsilon = [0.5, ..., 4]$ and in [239] the range is $\epsilon = [0.1, ..., 10]$).

Because our estimators in Eq. (7.1), Eq. (7.3), and Eq. (7.5) are unbiased, their variance is equal to the MSE (cf. Eq. (2.6)), which is commonly used in practice as an accuracy metric [202, 203, 239, 224]. So, to evaluate our results, we use the MSE metric averaged per the number of attributes d to evaluate our results. Thus, for each attribute j, we compute for each value $v_i \in A_j$ the estimated frequency $\hat{f}(v_i)$ and the real one $f(v_i)$ and calculate their differences. More precisely,

$$MSE_{avg} = \frac{1}{d} \sum_{j \in [1,d]} \frac{1}{|A_j|} \sum_{v \in A_j} (f(v_i) - \hat{f}(v_i))^2. \tag{7.8}$$

Methods evaluated. We consider for evaluation the following solutions (cf. Fig. 7.1) and protocols:

- Solution *Spl*, which splits the privacy budget per attribute ϵ/d with a best-effort approach using the adaptive mechanism presented in Section 2.4.4, i.e., Spl[ADP].
- Solution *Smp*, which randomly samples a single attribute and use all the privacy budget ϵ also with the adaptive mechanism, i.e., Smp[ADP].
- Our solution RS+FD, which randomly samples a single attribute and uses an amplified privacy budget $\epsilon' \geq \epsilon$ while generating fake data for each $d-1$ non-sampled attribute:
  - RS+FD[GRR] (Alg. 4 with GRR as local randomizer $\mathcal{A}$);
  - RS+FD[OUE-z] (Alg. 5);
  - RS+FD[OUE-r] (Alg. 6);
  - RS+FD[ADP] presented in Section 7.2.4 (i.e., adaptive choice between RS+FD[GRR] and RS+FD[OUE-z]).

7.3.2/ RESULTS ON SYNTHETIC DATA

Our first set of experiments were conducted on six synthetic datasets. Fig. 7.6 (first two synthetic datsets), Fig. 7.7 (third and fourth synthetic datsets), and Fig. 7.8 (last two synthetic datasets) illustrate for all methods, the averaged MSE_{avg} (y-axis) according to the privacy parameter ϵ (x-axis).

Impact of the number of users. In both Fig. 7.6 and Fig. 7.7, one can notice that the MSE_{avg} decreases with respect to the number of users n. More precisely, with the datasets we experimented, the MSE_{avg} decreases (approximately) one order of magnitude by increasing n in one order of magnitude too. In comparison with *Smp*, the noise in our RS+FD solution comes mainly from fake data as it uses an amplified $\epsilon' \geq \epsilon$. *This*

suggests that, in some cases, with appropriately high number of user n, our solutions may most likely provide higher data utility than the state-of-the-art Smp solution (e.g., cf. Fig. 7.8).

Impact of the number of attributes. One can notice the effect on increasing d comparing the results of Fig. 7.6 ($d = 5$) and Fig. 7.7 ($d = 10$) while fixing n and **c** (uniform number of values). For instance, even though there are twice the number of attributes, the accuracy (measured with the averaged MSE metric) does not suffer much. *This is because the amplification by sampling ($\frac{e^{\epsilon'}-1}{e^{\epsilon}-1} = \frac{1}{\beta}$ [43]) depends on the sampling rate $\beta = \frac{1}{d}$, which means that the more attributes one collects, the more the ϵ' is amplified, i.e., $\epsilon' = \ln(d \cdot (e^{\epsilon} - 1) + 1)$; thus balancing data utility.*

Besides, in Fig. 7.8, one can notice a similar pattern, i.e., increasing the number of attributes from $d = 10$ (left-side plot) to $d = 20$ (right-hand plot), with varied domain size **c**, resulted in only a slightly loss of performance. This, however, is not true for the *Spl* solution, for example, in which the MSE_{avg} increased much more in order of magnitude than our RS+FD solution.

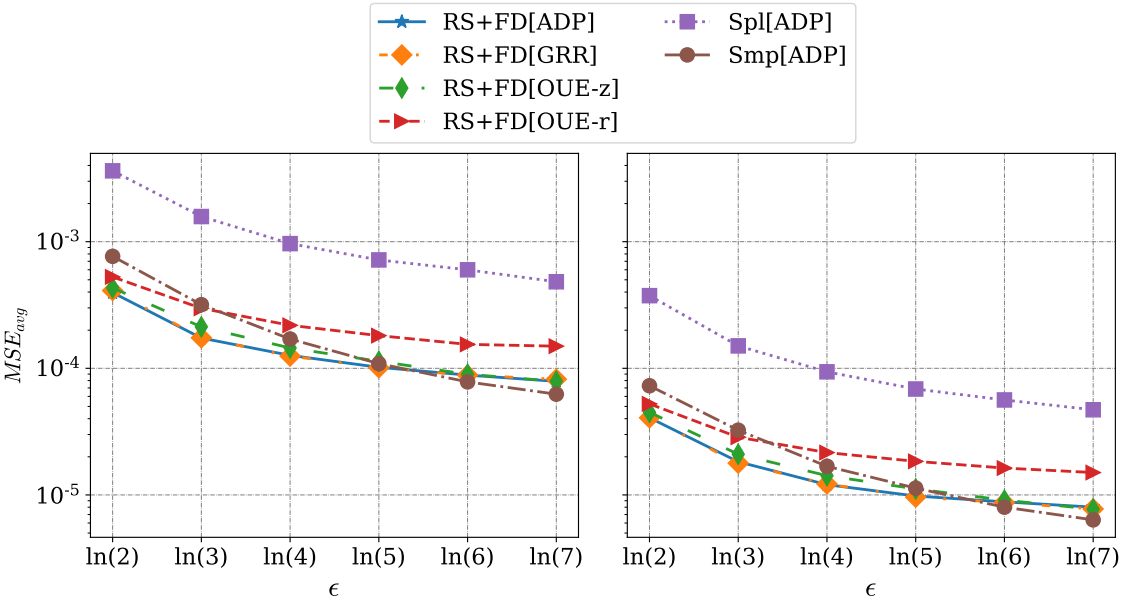

Figure 7.6: Averaged MSE varying ϵ on the *synthetic* datasets with $d = 5$, uniform domain size $\mathbf{c} = [10, 10, ..., 10]$, and $n = 50000$ (left-side plot) and $n = 500000$ (right-side plot).

Comparison with existing solutions. From our experiments, one can notice that the *Spl* solution always resulted in more estimation error (i.e., higher MSE_{avg}) than our RS+FD solution and than the *Smp* solution, which is in accordance with other works [166, 83, 108, 215, 239]. Besides, our RS+FD[GRR], RS+FD[OUE-z], and RS+FD[ADP] protocols achieve smaller estimation error (i.e., lower MSE_{avg}) or nearly the same MSE_{avg} than the *Smp* solution with a best-effort adaptive mechanism Smp[ADP], which uses GRR for small domain sizes k and OUE for large ones. Although this is not true with RS+FD[OUE-

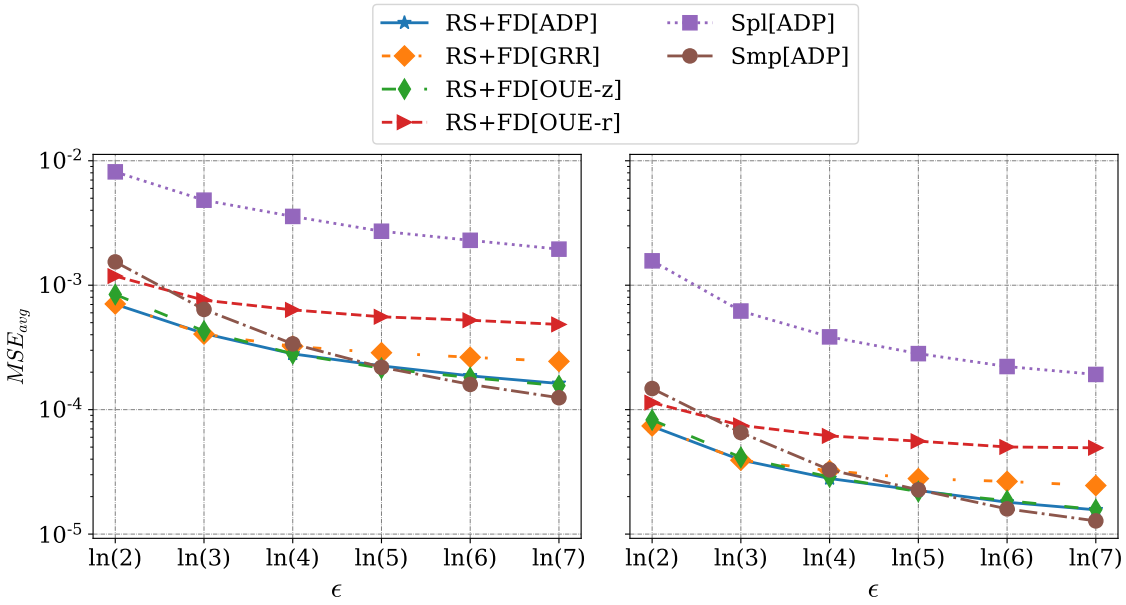

Figure 7.7: Averaged MSE varying ϵ on the *synthetic* datasets with $d = 10$, uniform domain size $\mathbf{c} = [10, 10, ..., 10]$, and $n = 50000$ (left-side plot) and $n = 500000$ (right-side plot).

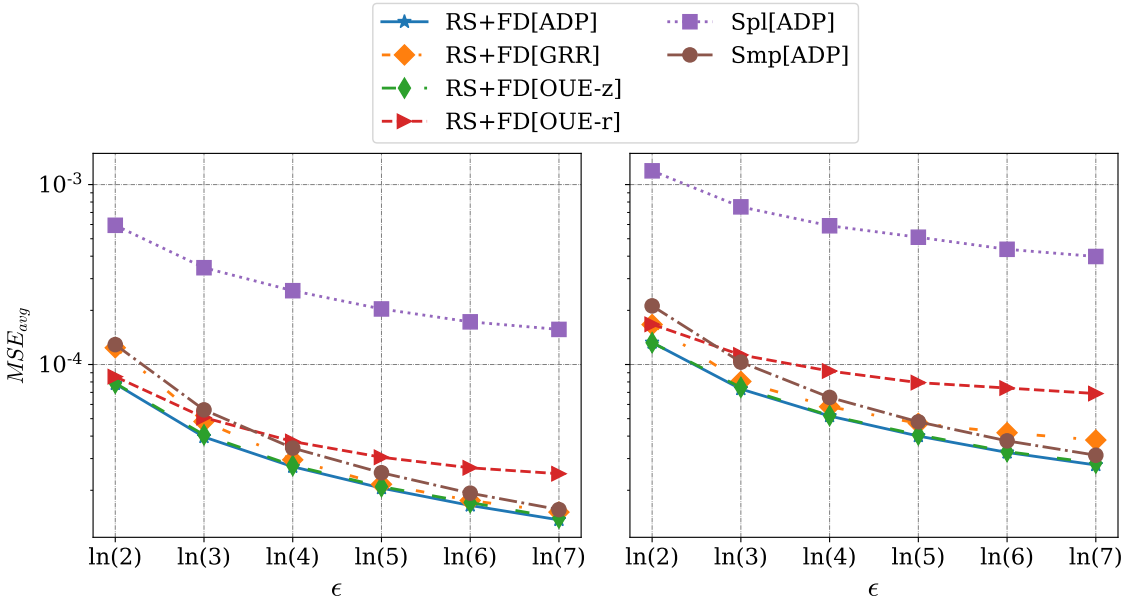

Figure 7.8: Averaged MSE varying ϵ on the *synthetic* datasets with $n = 500000$: the first with $d = 10$ and domain size $\mathbf{c} = [10, 20, ..., 90, 100]$ (left-side plot), and the other with $d = 20$ and domain size $\mathbf{c} = [10, 10, 20, ..., 100, 100]$ (right-side plot).

r], it still provides less estimation error than Spl[ADP] while "hiding" the sampled attribute from the aggregator.

Globally, on high privacy regimes (i.e., low values of ϵ), our RS+FD solution consistently outperforms the other two solutions *Spl* and *Smp*. By increasing ϵ,

Smp[ADP] starts to outperform RS+FD[OUE-r] while achieving similar performance than our RS+FD[GRR], RS+FD[OUE-z], and RS+FD[ADP] solutions. In addition, one can notice in Fig. 7.7, for example, the advantage of RS+FD[ADP] over our protocols RS+FD[GRR] and RS+FD[OUE-z] applied individually, as it adaptively selects the protocol with the smallest *approximate variance* value.

7.3.3/ RESULTS ON REAL WORLD DATA

Our second set of experiments were conducted on four real-world datasets with varied parameters for n, d, and **c**. Fig. 7.9 (*Nursery*), Fig. 7.10 (*Adult*), Fig. 7.11 (*MS-FIMU*), and Fig. 7.12 (*Census-Income*) illustrate for all methods, averaged MSE_{avg} (y-axis) according to the privacy parameter ϵ (x-axis).

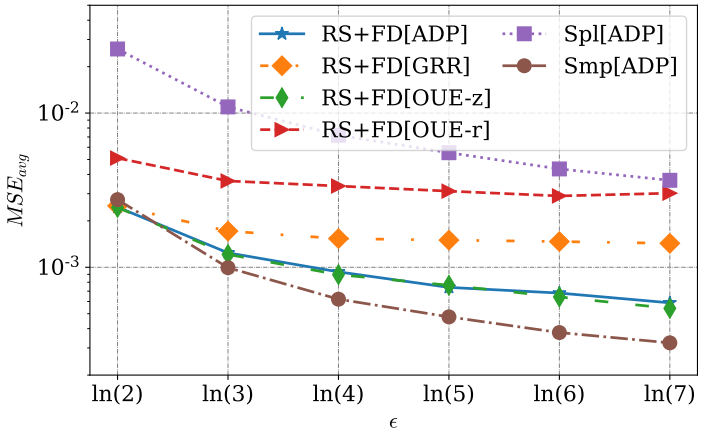

Figure 7.9: Averaged MSE varying ϵ on the *Nursery* dataset with $n = 12960$, $d = 9$, and domain size $\mathbf{c} = [3, 5, 4, 4, 3, 2, 3, 3, 5]$.

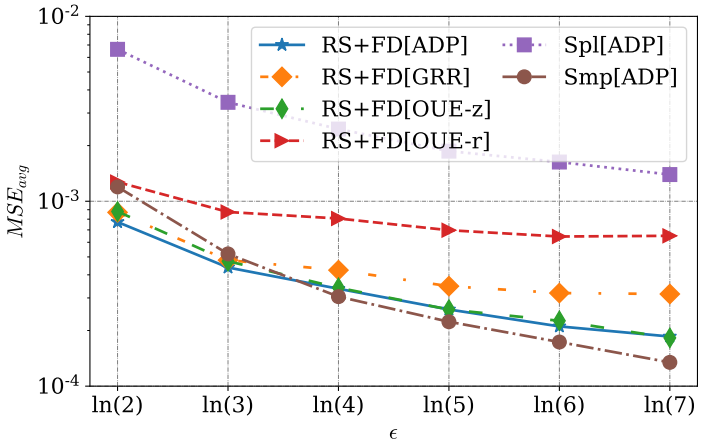

Figure 7.10: Averaged MSE varying ϵ on the *Adult* dataset with $n = 45222$, $d = 9$, and domain size $\mathbf{c} = [7, 16, 7, 14, 6, 5, 2, 41, 2]$.

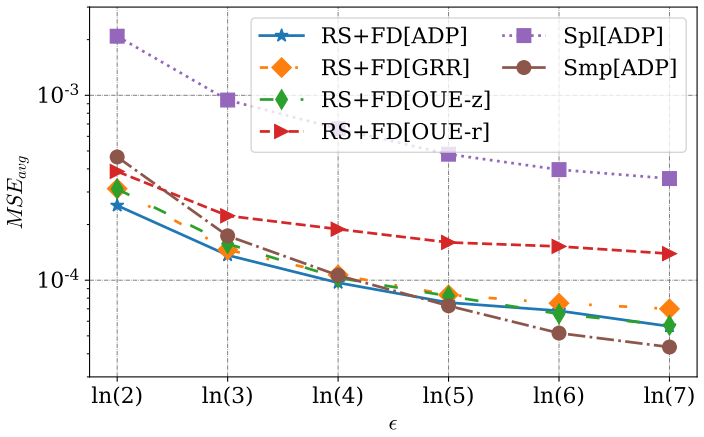

Figure 7.11: Averaged MSE varying ϵ on the *MS-FIMU* dataset with $n = 88935$, $d = 6$, and domain size $\mathbf{c} = [3, 3, 8, 12, 37, 11]$.

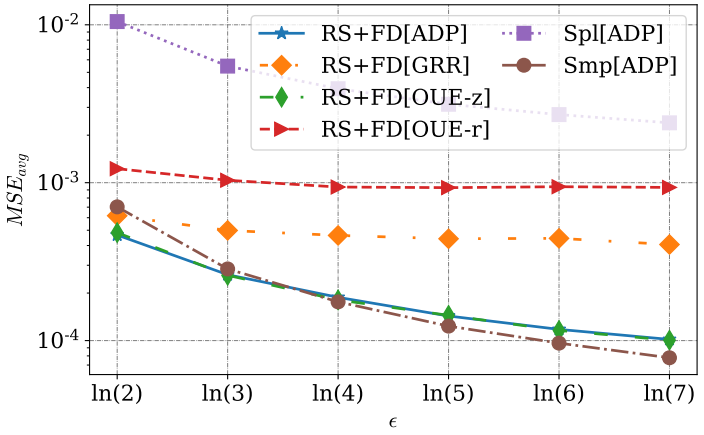

Figure 7.12: Averaged MSE varying ϵ on the *Census-Income* dataset with $n = 299285$, $d = 33$, and domain size $\mathbf{c} = [9, 52, 47, 17, 3, ..., 43, 43, 43, 5, 3, 3, 3, 2]$.

The results with real-world datasets follow similar behavior than with synthetic ones. For all tested datasets, one can observe that the MSE_{avg} of our proposed protocols with RS+FD is still smaller than the *Spl* solution with a best-effort adaptive mechanism Spl[ADP]. As also highlighted in the literature [166, 83, 108, 215, 239] and in Chapters 5 and 6, privacy budget splitting is sub-optimal, which leads to higher estimation error.

On the other hand, for both *Adult* and *MS-FIMU* datasets, our solutions RS+FD[GRR], RS+FD[OUE-z], and RS+FD[ADP] achieve nearly the same MSE_{avg} (sometimes smaller MSE_{avg} on high privacy regimes, i.e., for low ϵ) than the *Smp* solution with the best-effort adaptive mechanism Smp[ADP]. For the *Nursery* dataset, with small number of users n, only RS+FD[OUE-z] and RS+FD[ADP] are competitive with Smp[ADP]. Lastly, for the *Census* dataset, with a large number of attributes $d = 33$, increasing the privacy parameter ϵ resulted in a small gain on data utility for our solutions RS+FD[GRR] and RS+FD[OUE-r]. On the other hand, both of our solutions RS+FD[OUE-z] and

RS+FD[ADP] achieve nearly the same or smaller MSE_{avg} scores than Smp[ADP].

Moreover, one can notice that using the *approximate variance* in Eq. (7.7) led RS+FD[ADP] to achieve an improved performance over our RS+FD[GRR] and RS+FD[OUE-z] protocols applied individually. For instance, for the *Adult* dataset, with RS+FD[ADP] it was possible to outperform Smp[ADP] 3x more than with RS+FD[GRR] or RS+FD[OUE-z] (similarly, 1x more for the *MS-FIMU* dataset). Besides, for the *Census-Income* dataset, RS+FD[ADP] improves the performance of the other protocols applied individually on high privacy regimes while accompanying the RS+FD[OUE-z] curve on the lower privacy regime cases.

In general, these results help us answering the problematic of this chapter (cf. Section 7.1) that for the same privacy parameter ϵ, one can achieve nearly the same or better data utility with our RS+FD solution than when using the state-of-the-art *Smp* solution. Besides, RS+FD enhances users' privacy by "hiding" the sampled attribute and its ϵ-LDP value among fake data. On the other hand, there is a price to pay on computation, in the generation of fake data, and on communication cost, which is similar to the *Spl* solution, i.e., send a value per attribute.

7.4/ Discussion and Related Work

As reviewed in Section 6.4.3, most studies for collecting multidimensional data with LDP mainly focused on numerical data [83, 123, 166, 239] or other complex tasks with categorical data, e.g., marginal estimation [233, 161, 138, 134, 78] and analytical/range queries [208, 207, 153, 147]. Regarding multidimensional frequency estimates, in Chapters 5 and 6, we prove that for GRR, SUE, and OUE, sending a single attribute with the whole privacy budget ϵ results in less variance than splitting the privacy budget for all attributes, which is a common result in LDP literature [108, 208, 238, 184, 88].

However, in the aforementioned works [83, 166, 108, 123, 239] as well as in Chapters 5 and 6, the sampling result is known by the aggregator. That is, each user samples a single attribute j, applies a local randomizer to v_j, and sends to the aggregator the tuple $y = \langle j, LDP(v_j)\rangle$ (i.e., *Smp*). While one can achieve higher data utility (cf. Figs. 7.6- 7.12) with *Smp* than splitting the privacy budget among d attributes (*Spl*), we argue that *Smp* might be "unfair" with some users. More precisely, users whose sampled attribute is socially "more" sensitive (e.g., disease or location), might hesitate to share their data as the probability bound e^{ϵ} is "less" restrictive than $e^{\epsilon/d}$. For instance, assume that GRR is used with k=2 (HIV positive or negative) and the privacy budget is $\epsilon = ln(7) \sim 2$, the user will report the true value with probability as high as $p \sim 87\%$ (even with $\epsilon = 1$, this probability is still high $p \sim 73\%$). On the other hand, if there are $d = 10$ attributes (e.g.,

nine demographic and HIV test), with *Spl*, the probability bound is now $e^{\epsilon/10}$ and $p \sim 55\%$.

Motivated by this privacy-utility trade-off between the solutions *Spl* and *Smp*, we proposed a solution named random sampling plus fake data (RS+FD), which generates uncertainty over the sampled attribute in the view of the aggregator. In this context, since the sampling step randomly selects an attribute with sampling probability $\beta = \frac{1}{d}$, there is an amplification effect in terms of privacy, a.k.a. amplification by sampling [25, 43, 118, 173, 32]. A similar privacy amplification for sampling a random item of a single attribute has been noticed in [136] for frequent itemset mining in the LDP model too. Indeed, *amplification* is an active research field on DP literature, which aims at finding ways to measure the privacy introduced by non-compositional sources of randomness, e.g., sampling [25, 43, 118, 173, 32], iteration [125], and shuffling [141, 149, 184, 202, 224].

7.5/ CONCLUSION

In this chapter, we proposed a solution, namely, RS+FD for multidimensional frequency estimates under ϵ-LDP, which is generic to be used with any existing LDP mechanism developed for single-frequency estimation. More precisely, with RS+FD, the client-side has two steps: local randomization and fake data generation (cf. Fig. 7.1 and Alg. 4). First, an LDP mechanism preserves privacy for the data of the sampled attribute. Second, the fake data generator provides fake data for each $d-1$ non-sampled attribute. This way, the sanitized data is "hidden" among fake data and, hence, the sampling result is not disclosed along with the users' report (and statistics).

What is more, we notice that RS+FD can enjoy privacy amplification by sampling [25, 43, 118, 173, 32], detailed in Section 2.3.3. That is, if one randomly sample a dataset without replacement using a sampling rate $\beta < 1$, it suffices to use a privacy budget $\epsilon' \geq \epsilon$ to satisfy ϵ-DP, where $\frac{e^{\epsilon'}-1}{e^{\epsilon}-1} = \frac{1}{\beta}$ [43]. This way, given that the sampled dataset for each attribute has non-overlapping users, i.e., each user selects an attribute with sampling probability $\beta = \frac{1}{d}$, to satisfy ϵ-LDP, each user can apply an LDP mechanism with $\epsilon' = \ln(d \cdot (e^{\epsilon} - 1) + 1) \geq \epsilon$.

Moreover, we integrated two state-of-the-art LDP mechanisms, namely, GRR [80] and OUE [108], within RS+FD to develop four protocols: RS+FD[GRR], RS+FD[OUE-z], RS+FD[OUE-r], and RS+FD[ADP]. We analyze these four protocols analytically and experimentally through a comprehensive and extensive set of experiments on both synthetic and real-world open datasets. With our experiments, we can conclude that **under the same privacy guarantee, our proposed protocols with RS+FD achieve similar or better utility (measured with the MSE_{avg} metric) than using the state-of-the-art *Smp* solution (see Figs. 7.6 – 7.12).** Besides these **utility results**, RS+FD also generates *uncertainty* over the sampled attribute in the view of the aggregator, which enhances

users' privacy.

IV

Contribution: Differentially Private Machine Learning Predictions

8

Forecasting Mobility Data With Differentially Private Deep Learning

In Chapters 5-7 we have focused and contributed on **statistical learning** with the local DP model. From this Chapter 8 until Chapter 11, we concentrate our efforts on **differentially private machine learning**. As mentioned in Chapter 1, we aim to solve real-world problems using machine learning, assuming centralized data owners (e.g., MNOs and EMS) that collect sensitive information from individuals for both billing and/or legal purposes. This way, we consider settings applying either centralized DP algorithms (Chapters 8 and 11) or LDP algorithms (Chapters 9 and 10) to sanitize the data on the server-side, which is ϵ-DP for users. However, besides sanitizing the data, extracting meaningful predictions is also of great interest, thus, requiring a proper evaluation of the privacy-utility trade-off.

Moreover, in Chapters 1 and 4, we have reviewed mobility reports published by OBS Flux Vision system [53] and in Chapter 5 we have proposed an LDP-based CDRs processing system as a stronger alternative to "anonymity on-the-fly", i.e., with "sanitization on-the-fly". **In this chapter, we assume that besides generating mobility reports, MNOs (or any involved entity) could also be interested in forecasting aggregate human mobility statistics.** Therefore, in this chapter, we will assume the existence of two settings for privacy-preserving human mobility analytics using CDRs. The first scenario, $\mathbf{S}_1$, considers that aggregated mobility statistics are published following the anonymity "on-the-fly" model of MNOs CDRs processing systems (e.g., as in [53]). The second setting, $\mathbf{S}_2$, considers that besides anonymity "on-the-fly", a centralized DP algorithm (e.g., Laplace or Gaussian mechanisms from Section 2.3) is used to sanitize the aggregate mobility statistics before public release.

In other words, this corresponds to evaluating the privacy-utility trade-off of applying centralized DP algorithms to the current anonymity-based statistics. This is the core contri-

bution of this chapter, in which we evaluate differentially private deep learning models for multivariate time-series forecasting of aggregate human mobility data. Notice that while the previous chapters considered multiple attributes, we will focus on a single attribute here, namely, the number of people per several given regions.

8.1/ INTRODUCTION

As reviewed in Chapters 1, 4, and 5, on analyzing mobility data, some studies have shown that humans follow particular patterns with predictability [49] and, hence, *users' privacy is a major concern* [49, 122, 52, 66, 55, 38, 56, 103, 198, 135, 109]. Because of these privacy issues, MNOs tend to publish aggregated mobility data [218, 109, 237, 135, 53], e.g., the number of users in given areas at a given timestamp, which, in other words, represents a **multivariate time series dataset**.

However, as recent studies have shown, even aggregated mobility data (e.g., heatmaps) can be subject to membership inference attacks [103, 198] and users' trajectory recovery attack [135, 109]. More precisely, the later authors in [135, 109] showed that their attack reaches accuracies as high as $73\% \sim 91\%$, suggesting generalization and perturbation through DP [27, 26, 59] as a means to mitigate this attack.

With these elements in mind, this chapter contributes with a comparative analysis between adding DP guarantees into two different steps of training deep learning (DL) models to **forecasting multivariate aggregated human mobility data**. On the one hand, we consider using **gradient perturbation**, which can be achieved by training DL models over original time-series data with the DP-SGD [73, 131, 241] algorithm. This case corresponds to collecting data following the scenario $\mathbf{S}_1$ mentioned at the beginning of this chapter and training a differentially private DL model. On the other hand, we consider using **input data perturbation**, i.e., training DL models with differentially private time series data. This corresponds to collecting data following the scenario $\mathbf{S}_2$ also mentioned at the beginning of this chapter and training any non-private DL model on it. We have briefly presented both gradient and input perturbation settings in Section 3.2.

We carried out our experiments with the real-world mobility dataset collected by OBS [53] named Paris-DB described in Section 3.3.1.2. In this chapter, **we aim at forecasting the future number of people at the next 30-min interval in each of the 6 regions**. That is, given $X_{(t_1,t_\tau)}$, the goal is to forecast $X_{(t_{\tau+1})}$, i.e., **one-step-ahead forecasting**, which is unknown at time τ. Therefore, we benchmark four state-of-the-art DL models (i.e., recurrent neural networks) with the Paris-DB, providing a first comparative evaluation on the impact of differential privacy guarantees when training DL models in both input and gradient perturbation settings. Indeed, we intend that from this study, other

classical multivariate time series forecasting, ML, and privacy-preserving ML techniques can be tested and compared. We invite the interested reader to also visit the **Github page (https://github.com/hharcolezi/ldp-protocols-mobility-cdrs), in which we release the dataset and codes we used for our experiments.**

The remainder of this chapter is organized as follows. In Section 8.2, we present the experimental setup, our results and its discussion, and we review related work. Lastly, in Section 8.3, we present the concluding remarks and future directions. The experiments and results in Sections 8.2 and 8.3 were submitted as part of a full article to the Neural Computing and Applications journal.

8.2/ EXPERIMENTAL VALIDATION

We divide this section in the following way. First, we describe general settings for our experiments (Section 8.2.1). Next, we present the development and evaluation of non-private DL models (Section 8.2.2). Lastly, we present the development of differentially private DL models, which include both gradient and input perturbation settings (Section 8.2.3).

8.2.1/ GENERAL SETUP OF EXPERIMENTS

Environment. All algorithms were implemented in Python 3.8.8 with Keras [68] and Tensorflow Privacy (TFP) [131] libraries.

Dataset. In this chapter, we only utilize the second period of the Paris-DB from Section 3.3.1.2, which has aggregated mobility data for 72 days (from 2020-08-24 to 2020-11-04). We split the Paris-DB into exclusively divided learning (first 65 days, i.e., $n_l = 3120$ intervals of 30-min) and testing (last 7 days, i.e., $n_t = 336$ intervals of 30-min) sets. Table 8.1 presents descriptive statistics about both dataset with the following measures per region (labeled as R1 - R6): min, max, mean, standard deviation (std), and median.. Fig. 8.1 exemplifies the data separation into train and test sets for region R1.

Table 8.1: Descriptive statistics for the multivariate time series dataset on the number of users per coarse region.

Statistic	R1	R2	R3	R4	R5	R6
Min	56,937	1,996	1,429	255	252	347
Max	165,405	21,980	28,990	25,184	7,961	27,637
Mean	116,777	14,307	16,274	11,758	4,166	11,559
Std	17,947	2,803	3,915	3,682	1,450	5,136
Median	121,488	14,808	16,661	12,134	4,495	12,542

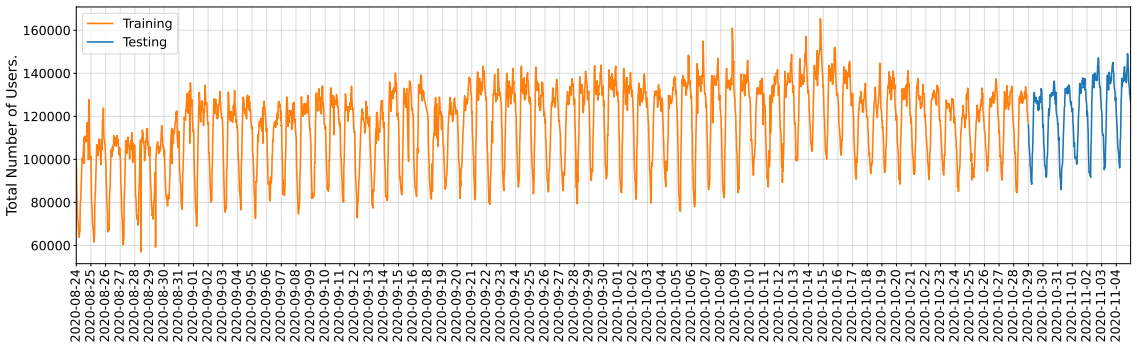

Figure 8.1: Example of data separation into training and testing sets for region R1.

Temporal features. We added the time of the day and the time of the week as cyclical features to help models recognizing low and high peak values of human mobility patterns.

Forecasting methodology. We used 6 prior time steps (i.e., lag values), which showed autocorrelation higher than 0.5 to predict a single step ahead in the future (i.e., short forecasting horizon). More specifically, the forecasting models will take into account the number of people in each region from 3 hours to make predictions *one-step-ahead* for each region in the next 30-min interval. And in the end, we compute the performance metrics.

Performance metrics. All models were evaluated with standard time-series metrics, namely, RMSE and MAE, both explained in Section 3.1.5. RMSE was the primary metric to select the final DL models. As a multi-output scenario (i.e., 6 regions), we present the metrics per region as well as its averaged values. In all experiments, due to randomness, we report the results of the model with the lowest RMSE over 10 runs.

8.2.2/ NON-PRIVATE DL FORECASTING MODELS

Baseline model. We established a naive forecasting technique a.k.a. "persistence model", which for each region, it returns the current number of people at time t as the forecasted value, i.e., $\mathbf{x}_{t+1} = \mathbf{x}_t$. Notice that this is a quite accurate baseline since, in general, the number of people per region varies slowly by 30-min (i.e., walking people may take more time to move from one area to another).

Methods evaluated. To predict the number of users in each region in a multivariate time series forecasting framework, we compared the performance of four state-of-the-art DL models, i.e., recurrent neural networks: LSTM [16], GRU [57], and their Bidirectional [17] architectures, i.e., BiLSTM and BiGRU. These methods have been briefly presented in Section 3.1.3.3.

Model selection. To optimize the hyperparameters per DL method, we used Bayesian optimization [47] (explained in Section 3.1.6) with 100 iterations to minimize $loss =$

$RMSE_{avg} + RMSE_{std}$; the subscripts *avg* and *std* indicates the averaged and standard deviation values of the RMSE metric considering the 6 regions. For each method, we only used a single hidden layer followed by a dense layer (output), since RNNs generally perform well with a low number of hidden layers [222]. So, we searched the following hyperparameters: number of neurons (h_1), batch size (bs), and learning rate (η). All models used "relu" (rectified linear unit) as activation function, which resulted in better performance than the default "tanh" activation in prior tests. Lastly, models were trained using the adam (adaptive moment estimation) optimizer during 100 epochs by minimizing the MAE loss function. Table 8.2 exhibits the hyperparameters' search space and the final value used per DL method.

Table 8.2: Search space for hyperparameters and the final configuration obtained by DL method.

Hyperparameter's range	Step	LSTM	BiLSTM	GRU	BiGRU
h_1: [25 – 500]	25	225	500	75	175
bs: [5 – 40]	5	10	10	5	5
η: [1e-5 – 3e-3]	–	0.002233	0.002303	0.001725	0.000289

Results and analysis. Table 8.3 present the performance of the developed DL models in comparison with the Baseline model based on RMSE and MAE metrics per region and the resulting mean. *Notice that the metrics are in the real scale according to the number of users per region (cf. Table 8.1). That said, although R1 presents higher metric values, it does not necessarily mean worse results.* One solution could be normalizing the data. Besides, Fig. 8.2 illustrates for each region forecasting results for the last day of our testing set, which includes the real number of people and the predicted ones by each RNN: LSTM, GRU, BiLSTM, and BiGRU.

Table 8.3: Performance of the Baseline model and non-private DL models based on RMSE and MAE metrics per region and the resulting mean values.

Model	Metric	R1	R2	R3	R4	R5	R6	Mean
Baseline	RMSE	3461.6	1131.8	1517.9	986.5	561.3	1362.3	1503.6
	MAE	2597.5	839.4	1105.8	744.1	434.3	921.5	1107.1
LSTM	RMSE	2667.2	1007.3	1291.6	887.2	536.3	1135.6	1254.2
	MAE	2053.8	758.1	969.8	662.6	432.3	786.0	943.8
BiLSTM	RMSE	2572.7	1033.3	1276.4	872.7	528.1	1166.7	1241.6
	MAE	1954.7	781.5	965.5	660.8	419.4	808.2	931.7
GRU	RMSE	2539.1	973.0	1296.0	953.5	499.9	1185.1	1241.1
	MAE	1949.7	722.8	939.6	740.2	396.4	829.1	929.6
BiGRU	RMSE	2560.3	968.3	1282.6	832.1	478.9	1163.7	**1214.3**
	MAE	1957.2	717.0	955.3	623.0	382.7	807.5	**907.1**

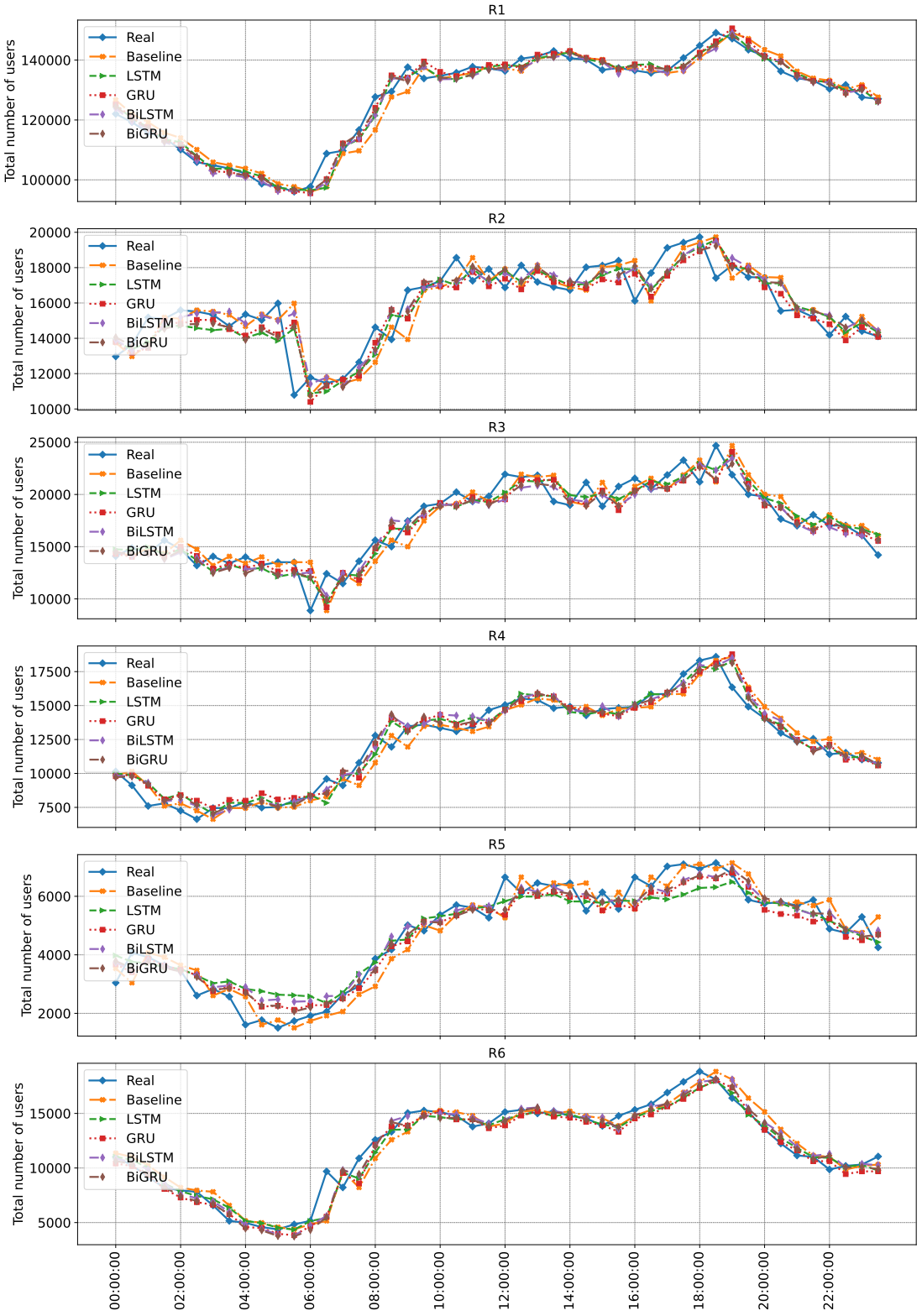

Figure 8.2: Multivariate time series forecast for the last day of the test set for the number of users per coarse region (R1 – R6) by the following models: Baseline, LSTM, GRU, BiLSTM, and BiGRU.

As one can notice, all DL models consistently outperform the Baseline model. On average, the BiGRU model outperformed all other forecasting methods, with results highlighted **in bold**. Indeed, for each region, the BiGRU consistently and considerably outperformed the Baseline model, showing the worthiness of developing DL models for this multivariate forecasting task. Similar scores were achieved by the GRU and BiLSTM models with an average RMSE around 1241. The least performing DL method in our dataset was the LSTM model. Extending the architectures, hyperparameters range, lag values (i.e., test with less or more input time steps), dropout layers, for example, could probably improve our models and change the resulting most performing technique. However, we will focus our attention on a comparative analysis of privacy-preserving DL methods in the next subsection and, thus, these possible extensions are left as future work.

8.2.3/ PRIVACY-PRESERVING DL FORECASTING MODELS

Methods evaluated. We consider two privacy-preserving ML settings presented in Section 2.3, namely, input perturbation (IP) and gradient perturbation (GP). Thus, we selected only the DL method that presented the smallest RMSE with original data, i.e., BiGRU (cf. Table 8.3). We will use BiGRU[IP] and BiGRU[GP] to indicate a BiGRU trained under input and gradient perturbation, respectively.

For the model selection stage, **we first start with BiGRU[GP] since it allows defining a range of ϵ, which is dependent on several hyperparameters of DP-SGD. For a fair comparison between both settings, we utilize the given range of ϵ to develop BiGRU[IP] models too.** Notice, however, that in both scenarios, (ϵ, δ)-DP can be ensured to each time series data sample. On the other hand, this also means that the same user may have contributed to all $n_l = 3120$ training samples and, thus, in the worst case, the sequential composition theorem [59] applies. With these elements in mind, we considered high privacy regimes ($\epsilon \ll 1$) such that the maximum $\check{\epsilon} = \sum_{i=1}^{n_l} \epsilon_i$ is compatible with real-world DP deployed systems [219]. **This way, ϵ corresponds to the lower bound (the user appears in a single data point) and $\check{\epsilon}$ represents the upper bound (the user appears in all data points).**

BiGRU[GP] model selection. In addition to standard hyperparameters h_1, bs, and η (cf. Section 8.2.2), we also included the TFP hyperparameters in the Bayesian optimization with 100 iterations to minimize $loss = (RMSE_{avg} + RMSE_{std})e^{\epsilon}$; **the multiplicative factor e^{ϵ} is a penalization on high values of ϵ, which varies depending on the hyperparameters used per iteration**. More specifically, given the number of training samples $n_l = 3120$, we fix the following hyperparameters: the number of epochs equal 100, $num\_microbatches = 5$, $noise\_multiplier$ equal $\{35, 70, 140, 500\}$, respectively, and $\delta = 10^{-7}$, which respects $\sum_{i=1}^{n_l} \delta_i < 1/n_l$ [59]. This way, we varied h_1, bs, η, and $l2\_norm\_clip$ ac-

cording to Table 8.4, which exhibits the hyperparameters' search space, the final value used per BiGRU[GP] model, **and the resulting privacy guarantee ϵ calculated with the `compute_dp_sgd_privacy` function** [131], and the overall $\check{\epsilon} = \sum_{i=1}^{n_l} \epsilon_i$. Lastly, all BiGRU[GP] models also used "relu" as activation function and were trained using the differentially private adam optimizer by minimizing the MAE loss function.

Table 8.4: Search space for standard and TFP hyperparameters, the final configuration per BiGRU[GP] model, the final privacy guarantee ϵ per time-series sample, and the maximum $\check{\epsilon}$ following the sequential composition theorem [59].

Hyperparameter	BiGRU[GP]$_1$	BiGRU[GP]$_2$	BiGRU[GP]$_3$	BiGRU[GP]$_4$
h_1: [25 – 500]	500	425	275	475
bs: [5 – 40]	5	5	10	5
η: [1e-5 – 3e-3]	0.002229	0.000455	0.000291	0.001235
$l2\_norm\_clip$: {1, 1.5, 2, 2.5}	2.5	2	1	2.5
$noise\_multiplier$: `fixed`	35	70	140	500
Privacy guarantee	$\epsilon_1 = 0.0650$	$\epsilon_2 = 0.0399$	$\epsilon_3 = 0.0357$	$\epsilon_4 = 0.0317$
	$\check{\epsilon}_1 = 202.8$	$\check{\epsilon}_2 = 124.488$	$\check{\epsilon}_3 = 111.384$	$\check{\epsilon}_4 = 98.904$

BiGRU[IP] model selection. We fix $\delta = 10^{-7}$ and we apply the Gaussian mechanism [59], by varying ϵ according to Table 8.4 (with their respective upper bound $\check{\epsilon}$), **to the whole time series data, as it would be done if such system had been deployed in real life. The metrics, however, are computed in comparison with original raw time series data**. Because input perturbation allows using any post-processing techniques, we used the same model selection methodology as for non-private BiGRU models to optimize the hyperparameters for BiGRU[IP] models. The resulting values per $\epsilon = [0.0650, 0.0399, 0.0357, 0.0317]$, respectively, are: $\text{BiGRU[IP]}_1 : \{h_1 = 200, bs = 5, \eta = 0.001993\}$, $\text{BiGRU[IP]}_2 : \{h_1 = 275, bs = 5, \eta = 0.001182\}$, $\text{BiGRU[IP]}_3 : \{h_1 = 200, bs = 10, \eta = 0.001333\}$, and $\text{BiGRU[IP]}_4 : \{h_1 = 200, bs = 10, \eta = 0.000842\}$.

Privacy-preserving results and analysis. Table 8.5 presents the performance of differentially private BiGRU models trained under input and gradient perturbation regarding the RMSE and MAE metrics per region and the resulting mean values. We also included in Table 8.5 the utility loss of differentially private BiGRU models in comparison with non-private ones, for both RMSE and MAE averaged metrics $\mathscr{E}$, calculated as:

$$\mathscr{U} = \frac{\mathscr{E}_{DP} - \mathscr{E}_{NP}}{\mathscr{E}_{NP}}, \tag{8.1}$$

in which $\mathscr{E}_{NP}$ is the result of Non-Private BiGRU (cf. averaged metric values **in bold** from Table 8.3) and $\mathscr{E}_{DP}$ refers to the results of either BiGRU[GP] or BiGRU[IP] models. Indeed, Eq. (8.1) will be positive unless the differentially private model outperforms the non-private one (which is not the case in our results).

We remarked in our experiments that since there is a sufficient number of users per time

Table 8.5: Performance of differentially private BiGRU models based on RMSE and MAE metrics per region and the resulting mean values. The last column $\mathscr{U}$ exhibits the utility loss of differentially private BiGRU models in comparison with non-private ones, for both RMSE and MAE averaged metrics expressed in %.

$\epsilon, \check{\epsilon}$ values	Model	Metric	R1	R2	R3	R4	R5	R6	Mean	$\mathscr{U}$
$\epsilon_1 = 0.0650$	BiGRU[GP]$_1$	RMSE	2561.4	1027.3	1254.7	866.7	498.5	1145.7	1225.7	0.9378
		MAE	1973.4	773.8	925.	644.1	397.9	781.5	916.0	0.9776
$\check{\epsilon}_1 = 202.8$	BiGRU[IP]$_1$	RMSE	2600.9	997.1	1304.0	852.7	483.8	1175.2	1235.6	1.7531
		MAE	1966.0	737.5	957.1	645.4	385.1	821.1	918.7	1.2753
$\epsilon_2 = 0.0399$	BiGRU[GP]$_2$	RMSE	2600.2	956.0	1268.5	841.5	515.0	1146.3	<u>1221.2</u>	<u>0.5672</u>
		MAE	1978.9	709.2	944.4	643.3	417.4	769.9	910.5	0.3713
$\check{\epsilon}_2 = 124.488$	BiGRU[IP]$_2$	RMSE	2592.2	978.4	1251.5	854.2	495.6	1158.6	<u>1221.8</u>	<u>0.6166</u>
		MAE	1986.1	737.1	910.9	653.9	393.2	813.9	915.9	0.9666
$\epsilon_3 = 0.0357$	BiGRU[GP]$_3$	RMSE	2580.5	990.0	1268.5	854.5	504.9	1154.3	**1225.5**	**0.9213**
		MAE	1938.8	753.0	942.8	659.7	406.3	773.6	912.4	0.5808
$\check{\epsilon}_3 = 111.384$	BiGRU[IP]$_3$	RMSE	2587.8	1004.7	1262.3	843.2	512.8	1186.2	**1232.9**	**1.5307**
		MAE	1963.1	755.8	957.5	636.9	414.6	811.8	923.3	1.7824
$\epsilon_4 = 0.0317$	BiGRU[GP]$_4$	RMSE	2560.8	978.3	1322.5	836.1	494.4	1195.4	1231.3	1.3990
		MAE	1956.2	715.1	989.2	633.6	392.0	821.6	917.9	1.1871
$\check{\epsilon}_4 = 98.904$	BiGRU[IP]$_4$	RMSE	2562.2	1012.2	1351.2	862.9	533.5	1168.8	1248.4	2.8072
		MAE	1955.6	756.8	1027.6	650.1	423.9	826.8	940.2	3.6454

series sample (cf. Table 8.1), it was still possible to make accurate forecasts in both privacy-preserving ML settings with the experimented range of (ϵ, δ)-DP. Indeed, from Table 8.5, one can notice that all differentially private BiGRU models achieved averaged RMSE lower than 1250, in which the worst result achieved by BiGRU[IP]$_4$ is just 2.8072% less precise than the non-private BiGRU model, comparing the utility metric for RMSE. What is more, in both gradient and input perturbation settings, differentially private BiGRU models achieved smaller error metrics than non-private LSTM, BiLSTM, and GRU models (cf. Table 8.3). For instance, both BiGRU[GP]$_2$ and BiGRU[IP]$_2$ reached similar scores in comparison with the non-private BiGRU model, with a utility loss of about 0.57% and 0.62% (for RMSE), respectively. These results are highlighted in <u>underlined</u> font, which represents our best results in terms of utility, with differentially private BiGRU models.

Interestingly, the accuracy (measured with the RMSE metric) of differentially private BiGRU models did not necessarily decrease according to more strict ϵ, i.e., lower values. One can note that results with ϵ_2 and ϵ_3 were more accurate than with ϵ_1. This way, in terms of a satisfactory *privacy-utility trade-off*, both BiGRU[GP]$_3$ (0.92% less accurate) and BiGRU[IP]$_3$ (1.53% less accurate) presented adequate metrics scores while satisfying a low value of ϵ, with results highlighted **in bold**. Indeed, in the worst-case scenario, a user that was present in each data point would have leaked $\check{\epsilon}_3 = 111.384$ at the end of 65 days (i.e., $\epsilon \sim 1.7$ per day), which follows real-world DP systems deployed by industry nowadays [232, 219].

The contribution of this research is significant for those involved in urban planning and decision-making [122], providing a solution to the human mobility multivariate forecast problem through RNNs and differentially private BiGRUs. In addition, we point out the research community to the Github page mentioned in the introduction section, in which we release the mobility dataset used in this paper for further experimentation with time series, machine learning, and privacy-preserving methods. The related literature to our work includes the generation of synthetic mobility data [133, 52, 172], the development of Markov models to infer travelers' activity pattern [137], and the development of privacy-preserving methods to analyze CDRs-based data [66, 38, 52, 55]. Besides, the work in [225] surveys non-private deep learning applications to mobility datasets in general. Concerning differentially private deep learning, one can find the application of gradient perturbation-based DL models for load forecasting [165], an evaluation of differentially private DL models in federated learning for health stream forecasting [189], the proposal of locally differentially private DL architectures [179], practical libraries for differentially private DL [131, 241], and theoretical research works [73, 71].

Lastly, Fig. 8.3 illustrates for each region forecasting results for the last day of our testing set, which includes the real number of people and the predicted ones by the following models: Baseline, non-private BiGRU, BiGRU[GP]$_3$, and BiGRU[IP]$_3$. As one can notice, similar forecasting results were achieved by both non-private and DP-based BiGRU models, which clearly outperforms the Baseline model. Lastly, between both input and gradient perturbation settings, BiGRU[GP] models took more time to execute than BiGRU[IP] models due to DP-SGD. In terms of accuracy, BiGRU[GP] models consistently outperformed BiGRU[IP] models for the same (ϵ, δ)-DP privacy level in our experiments. Nevertheless, BiGRU[GP] is trained over non-DP time-series data, which might be subject to, e.g., data leakage [228], membership inference attacks [103, 198], and users' trajectory recovery attacks [135, 109].

8.3/ Conclusion and Perspectives

In this chapter, we assumed the existence of two privacy-preserving MNOs CDRs processing system that collect and release multivariate aggregate human mobility data, described at the beginning of this chapter. However, along with collecting time-series data, extracting meaningful forecasts is also of great interest [128]. Thus, this chapter evaluated differentially private DL models in both input and gradient perturbation settings to forecast multivariate aggregated mobility time series data.

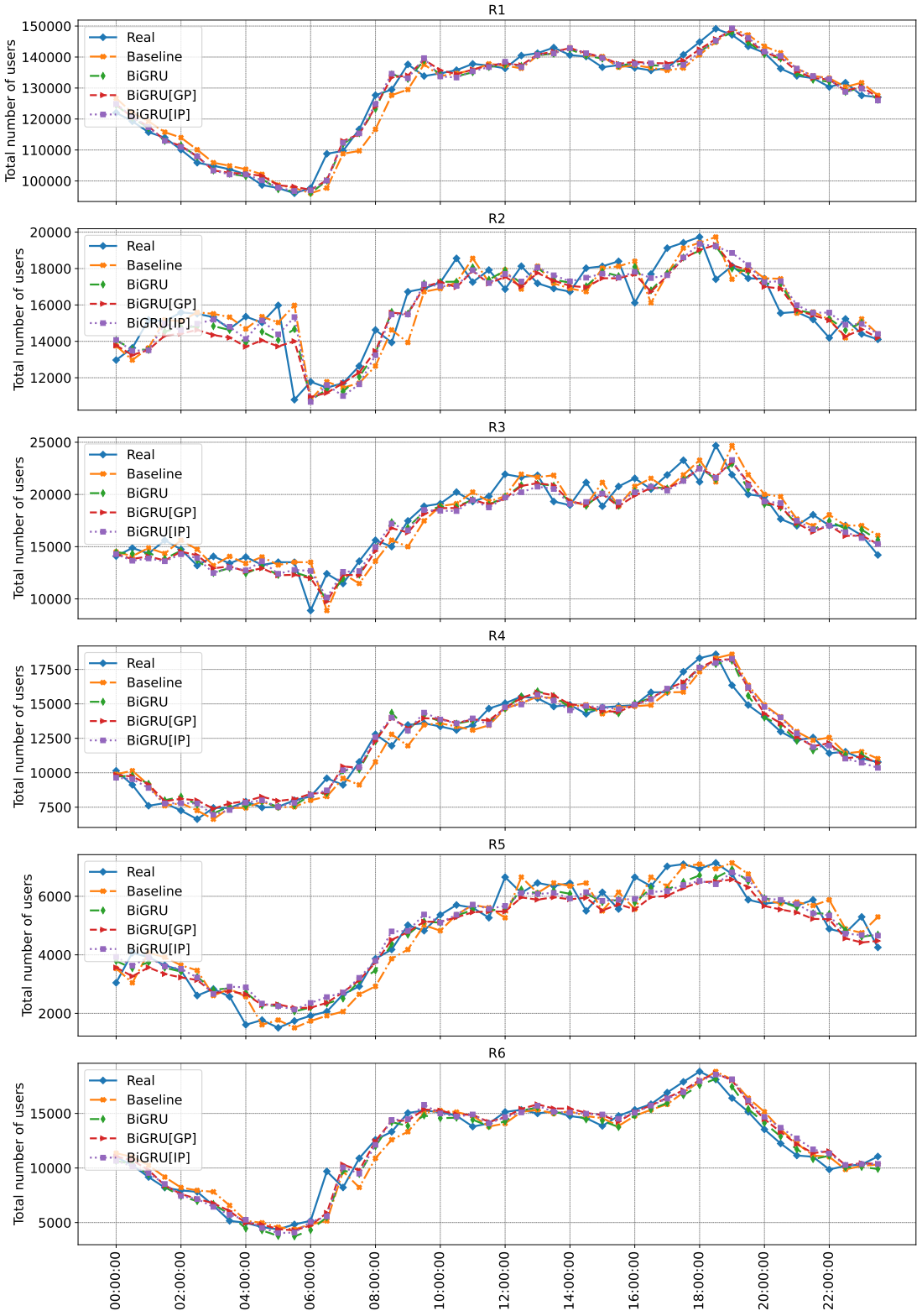

Figure 8.3: Multivariate time series forecast for the last day of the test set for the number of users per coarse region (R1 – R6) by the following models: Baseline, non-private BiGRU, BiGRU[GP]$_3$, and BiGRU[IP]$_3$.

Experiments were carried out on the dataset named Paris-DB from Section 3.3.1.2. First, we compared the performance of four non-private DL models (i.e., LSTM, GRU, BiLSTM, and BiGRU). Since the BiGRU model provided the highest utility, we selected it for building privacy-preserving models. Under gradient and input perturbation settings, i.e., BiGRU[GP] and BiGRU[IP], respectively, four values of $\epsilon \ll 1$ were evaluated. As shown in the results, differentially private BiGRU models achieve nearly the same performance as non-private BiGRU models, with utility loss related to the RMSE metric varying between $0.57\% - 2.8\%$.

Thus, we conclude that it is still possible to have accurate multivariate forecasts in both privacy-preserving ML settings, favoring the gradient perturbation setting in terms of accuracy and the input perturbation setting in terms of privacy protection. We believe that the input perturbation setting provides encouraging results for adding DP guarantees to MNOs CDRs processing systems, i.e., following the setting $\mathbf{S}_2$ mentioned at the beginning of this chapter. Indeed, besides being useful for forecasting tasks, DP would also add a layer of protection against, e.g., data breaches [228], membership inference attacks [103, 198], and users' trajectory recovery attacks [135, 109].

Some limitations and prospective directions of this chapter are described in the following. For differentially private BiGRU models, we only provided lower ϵ and upper $\check{\epsilon}$ bounds for the privacy guarantee of each sample in the time-series data. Using, however, advanced composition theorems [59] to account for the final privacy budget for each user was out of the scope of this chapter since the Paris-DB dataset does not contain users' IDs. Besides, although the developed DL models outperform the Baseline model ($\mathbf{x}_{t+1} = \mathbf{x}_t$), there is plenty of room for improvements to be carried out on hyperparameters optimization (e.g., accounting for the overall privacy budget [229]), data scaling, the number of lag values, etc. For instance, some high-peak values were missed by both non-private and DP-based DL models (see Fig. 8.3). In addition, we fixed the number of lagged values to 6 to predict a single step-ahead in the future (i.e., the forecasting horizon), in which the former can be tuned for performance improvement and the latter can be increased for multi-step forecasting tasks.

9

Forecasting Firemen Demand by Region With LDP-Based Data

In Chapter 8, we have started to evaluate the privacy-utility trade-off of differentially private machine learning models on a real-world problem concerning human mobility. From this Chapter 9 until Chapter 11, we will focus on our second motivating project (cf. Section 1.2), which concerns emergency medical services (EMS), in particular, using SDIS 25 [9] processed data by Selene Cerna. Similar to Chapter 8, this chapter also focuses on multivariate time-series forecasting but is related to the number of firefighters' interventions per region (referred to as firemen demand by region throughout this chapter). While there are several examples of EMS publicly sharing their data [8, 111, 158], we believe that more attention should be given to their victims' **privacy**. Indeed, the first question one may ask is if an intervention is a sensitive attribute. The answer is certainly yes because EMS would not have been called if the situation had not been severe enough (e.g., cardiac arrest, respiratory distress, ...). While the intention of the aforementioned EMS is laudable on publishing open-source data, there are many ways for misusing this information (e.g., discrimination in health insurance), which can jeopardize users' privacy.

Therefore, **in collaboration with Selene Cerna**, we propose in this chapter **a methodology based on generalization and LDP, which allows EMS to properly sanitize all their data row-by-row (i.e., independently). Thus, thanks to the post-processing properties of DP [59], EMS could use and/or share the sanitized data with third parties to develop ML-based decision-support tools**. Indeed, our solution envisaged both **statistical learning** and **machine learning** tasks (i.e., frequency estimation and multivariate time-series forecasting of firemen demand by region). We invite the reader to refer to Chapter 2 for the background on LDP. Lastly, similar to Chapter 5, we highlight that although we refer to our proposal as *LDP-based*, this is a centralizer data owner (i.e., EMS) that applies the LDP protocol on its servers, thus, providing ϵ-DP guarantees for users.

9.1/ INTRODUCTION

We start by recalling two recent publications of EMS French data on `data.gouv.fr`, a government site dedicated to initiatives in open data. The first concerns the 2007-2017 interventions of the Service Départemental d'Incendies et de Secours de Saône-et-Loire (SDIS 71), containing the number of interventions by type and by city [45][1]. The second concerns the same type of data for SDIS 91 (Essonne department) for the period 2010-2018 [111]. In each case, anonymization was done by aggregation: monthly for the first dataset and weekly for the second. Tables 9.1 and 9.2 exhibits five random samples of these datasets.

Table 9.1: Five random samples from the SDIS 71 dataset.

	Year_Month	ZIP Code	City	Aid to Person	Fire	...
1	2010-10	71093	LA CHAPELLE ST SAUVEUR	1	0	...
2	2012-04	71283	MARNAY	1	0	...
3	2012-11	71499	SANVIGNES LES MINES	11	2	...
4	2013-10	71221	GIVRY	10	3	...
5	2014-08	71017	BALLORE	1	0	...

Table 9.2: Five random samples from the SDIS 91 dataset (cf. [111]).

	Year	Week	ZIP Code	City	Category	Number of interventions
1	2018	24	91109	BRIERES-LES-SCELLES	Aid to Person	1
2	2018	39	91156	CHEPTAINVILLE	Aid to Person	1
3	2019	17	91215	EPINAY-SOUS-SENART	Fire in Urban Place	1
4	2019	28	91425	MONTLHERY	Aid to Person	10
5	2019	36	91629	VALPUISEAUX	Aid to Person	1

From Tables 9.1 and 9.2, one can notice that the applied anonymization method is both *too strong* and *too weak*. **Too strong**, first of all, because carrying out one aggregation per month results in the loss of useful information by summarizing the interventions at a cloud of 120 samples (12 per year), for which only a simple linear regression remains possible: it is hard to envisage machine learning with such a dataset - this is true, to a lesser extent, for data aggregated weekly. Then, **too weak**, because this aggregation per month, or per week, was done in a blind and generalized way: if some cities have a sufficiently large number of interventions, which allows a simple temporal aggregation to achieve anonymization of the data, others conversely do not have enough. In the case of monthly aggregation, for example, there are more than 600 situations where there has been only one intervention in a city in a given month: at this level, the simple 2-anonymity [18, 20] is no longer satisfied, and the information leakage is obvious. Such information leaks are also numerous in the case of weekly data, and **anonymization has failed for both sets of data**. For example, on analyzing the last row of Table 9.1, we

[1] Currently, this data is inaccessible on the referenced webpage.

learn, for example, that in the city of Ballore (FR-71220), an intervention took place in August 2014. Considering that the city has 86 inhabitants, it would not be very difficult to find the person who received help this month.

Moreover, a more risky case of possible data leakage concerns the Seattle Fire Department [8], which **displays live EMS response information with the precise hour, location, and reason for the incident**. While the intention of publishing precise EMS data is creditable, there are many ways for (mis)using this information, which can jeopardize users' privacy. For instance, if an attacker knows that one intervention took place in front of the house of a debilitated person, they may accurately infer that this person received care and (mis)use this information for their own good.

Therefore, the objective of this chapter is to propose a methodology that allows sanitizing each interventions' data with DP guarantees while still making it possible to aggregate them and make accurate forecasts. More specifically, **our proposal also utilizes generalization**, which, first, allows generalizing the precise hour to, e.g., days, and, secondly, to generalize the space of several small cities to bigger regions. Then, besides agglomeration of cities to regions, **we also propose that each intervention's region be ϵ-DP, which corresponds to applying an ϵ-LDP protocol row-by-row**. This way, EMS could share the sanitized dataset containing the generalized timestamp plus the ϵ-DP interventions' location data. **On the analyst side, one could, therefore, aggregate the sanitized data through frequency estimation (cf. Section 2.4) by different periods of interest**, e.g., daily, 4-days period, weekly, monthly, and so on, which would correspond to a **multivariate time-series dataset**.

In this chapter, we carried out our experiments with a real-world dataset collected by SDIS 25 [9] named **Interv-DB** from Section 3.3.3. The Interv-DB has information about 382046 **interventions** attended by the fire brigade from 2006 to 2018 inside their department. Our experiments are separated into two parts. The first concerns **statistical learning**, i.e., to estimate the frequency of firemen demand by region in different periods using the state-of-the-art OUE [108] protocol for single attribute frequency estimation (cf. Section 2.4). The second part is dedicated to **one-step-ahead forecasting** of firemen demand by region using daily estimated frequencies with the state-of-the-art XGBoost [76] ML technique.

The remainder of this chapter is organized as follows. Section 9.2 introduces our proposed methodology to sanitize EMS interventions' location data. Section 9.3 presents experiments on frequency estimation of firemen demand by region in different aggregation periods. Section 9.4 evaluates the privacy-utility trade-off of our methodology through training differentially private ML models over the sanitized multivariate time-series data. In Section 9.5, we conclude this work. A preliminary version of the proposed LDP-based methodology in Section 9.2 with its results was published in a full article [172] in the

Computers & Security journal.

9.2/ PROPOSED LDP-BASED METHODOLOGY

Fig. 9.1 illustrates the overview of the proposed privacy-preserving methodology that allows *EMS to sanitize their data*. We summarize our proposal in the following.

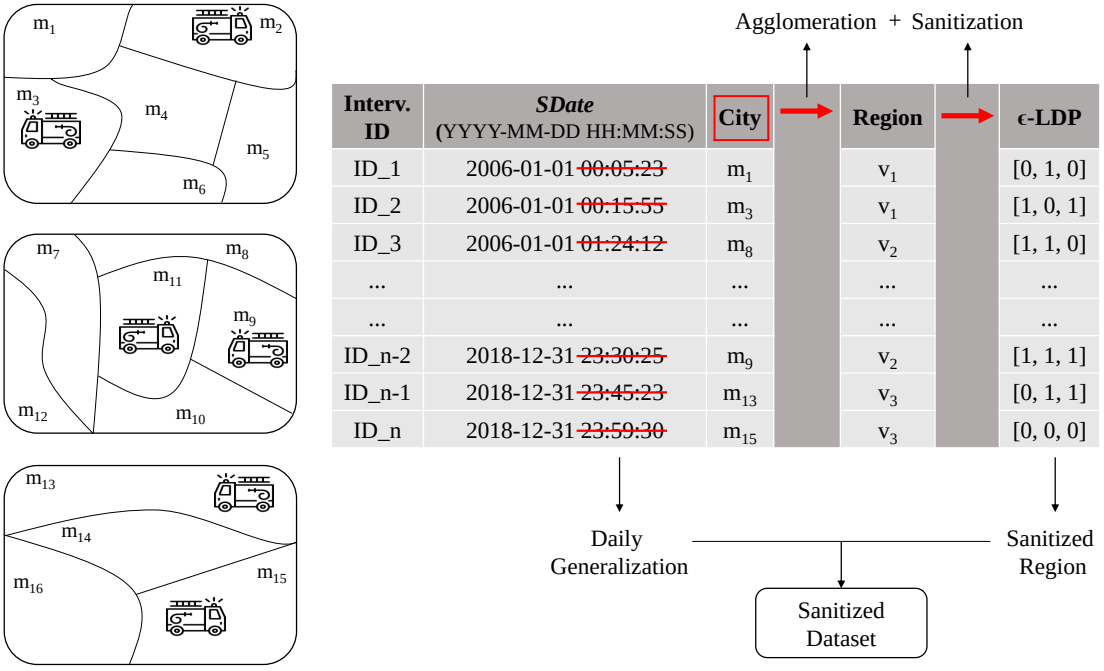

Figure 9.1: Overview of our LDP-based methodology to sanitize each firemen intervention independently.

Sanitization of Interventions' Data (EMS Side). From Fig. 9.1, the first step to guarantee the privacy of each interventions' data is through **generalization** of the starting date of the intervention (i.e., *SDate*) and the **agglomeration** of the location attribute. For the former, we adopt a **daily generalization** (not so strict as the datasets of Tables 9.2 and 9.1). For the latter, we propose a strong **agglomeration of small cities to bigger regions** in order to obtain events that are sufficiently representative in number. For example, one can notice in the left side of Fig. 9.1 that a set of $M = \{m_1, m_2, ..., m_{15}, m_{16}\}$ small cities are grouped to a set $A = \{v_1, v_2, v_3\}$ of 3 regions.

In this chapter, using the data at our disposal, 608 **cities where interventions happened in the Doubs department were generalized to** 17 **regions** using the public dataset available in [113]. The set $A = \{v_1, v_2, ..., v_{17}\}$ of 17 regions are: (1) CA du Grand Besançon, (2) CA Pays de Montbéliard Agglomeration, (3) CC Altitude 800, (4) CC de Montbenoit,

(5) CC des Deux Vallées Vertes, (6) CC des Lacs et Montagnes du Haut-Doubs, (7) CC des Portes du Haut-Doubs, (8) CC du Doubs Baumois, (9) CC du Grand Pontarlier, (10) CC du Pays d'Héricourt, (11) CC du Pays de Maîche, (12) CC du Pays de Sancey-Belleherbe, (13) CC du Plateau de Frasne et du Val Rasne et du Val de Drugeon (CFD), (14) CC du Plateau de Russey, (15) CC du Val de Morteau, (16) CC du Val Marnaysien, (17) CC Loue-Lison. Fig. 9.2 illustrates the department of Doubs with the respective cities and their agglomeration to regions.

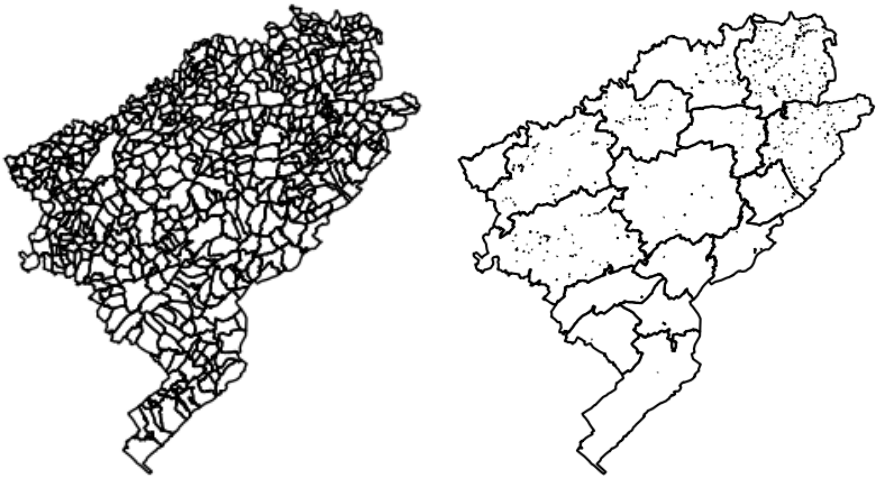

Figure 9.2: cities in the department of Doubs agglomerated by regions.

Up to now, the generalized dataset might still have unique events for a single day in a given region. However, **to improve the level of privacy for each intervention**, we propose to apply the OUE [108] LDP protocol row-by-row since it is independent on the number of regions. As presented in Section 2.4, given the original region $B = Encode(v)$ and the privacy budget ϵ, OUE reports a sanitized bit-vector B' where $\Pr[B'_i = 1] = p = \frac{1}{2}$ if $B_i = 1$ and $\Pr[B'_i = 1] = q = \frac{1}{e^{\epsilon+1}}$ if $B_i = 0$. Therefore, the final sanitized dataset (see the right side of Fig. 9.1) would be both generalized in terms of time and space and, besides that, it would be ϵ-DP, thus enhancing the privacy of each interventions' region.

Generating Synthetic Multivariate Time Series Datasets (Analyst Side). At this point, we assume that the analyst possess a sanitized dataset with generalized timestamp information and the ϵ-DP sanitized region (see the bottom right of Fig. 9.1). Since ϵ is a public parameter, **the analyst can define specific aggregation periods of their choice** (e.g., day, 3-days period, week, ...) and estimate the frequency $f(v_i)$ of firemen demand by region v_i with $\hat{f}(v_i) = \frac{N_i - nq}{n(p-q)}$, where N_i is the number of times the bit i has been reported and n is the number of interventions (cf. Eq. (2.4)). In this context, a synthetic **multivariate time-series dataset** will be built with all the estimated frequencies, which is considered as a non-interactive case of DP [110, 59]. Therefore, both *statistical learning* and *forecasting tasks* (following the input perturbation settings of Section 3.2.1) could be performed with

the aggregated dataset, while preserving privacy of the individuals concerned.

9.3/ Frequency Estimation of Firemen Demand by Region

9.3.1/ Setup of experiments

Environment. All algorithms were implemented in Python 3.8.8 with NumPy 1.19.5 and Numba 0.53.1 libraries. In all experiments, we report average results over 100 runs as LDP algorithms are randomized.

Dataset. We applied our proposed LDP-based methodology from Section 9.2 to the Interv-DB. The initial transformed dataset has both daily temporal information (the **when**) and the regions (the **where**) $382,046$ interventions took place from 2006 until 2018 (e.g., see the right-side of Fig. 9.1). This transformed **original dataset** will be used for frequency estimation experiments in this section and for forecasting tasks in the next Section 9.4.2.

Methods evaluated. In this chapter, we only applied the state-of-the-art OUE [108] LDP protocol for single attribute frequency estimation.

Evaluation and metrics. We considered **three different aggregation periods** in our experiments. The first aggregation period we analyze is with **yearly** data ($\tau = 13$ frequency estimates), which allows at the beginning of a year the fire brigade to better distribute their budget around its centers according to the firemen demand by region. Next, a **monthly** scenario ($\tau = 156$ frequency estimates) is considered. And, similar to before, the fire brigade can have high-utility statistics from a third-party company to reorganize budgets and personnel each month. Lastly, a **daily** scenario ($\tau = 4748$ frequency estimates) is taken into consideration such that ML algorithms could be applied in this amount of data.

For each aggregation period, similar to Chapter 7, we vary the privacy parameter in a logarithmic range as $\epsilon = [\ln(2), \ln(3), ..., \ln(7)]$. We use the MSE metric averaged by the number of aggregation periods τ to evaluate our results. Thus, for each time interval $t \in [1, \tau]$, we compute for each value $v_i \in A$ the estimated frequency $\hat{f}(v_i)$ and the original one $f(v_i)$ and calculate their differences. More precisely,

$$MSE_{avg} = \frac{1}{\tau} \sum_{t \in [1,\tau]} \frac{1}{|A|} \sum_{v_i \in A} (f(v_i) - \hat{f}(v_i))^2. \tag{9.1}$$

9.3.2/ FREQUENCY ESTIMATION RESULTS

Fig. 9.3 shows the relationship between MSE_{avg} and ϵ for all three aggregation periods, i.e., daily, monthly, and yearly. Moreover, for the sake of illustration, Fig. 9.4 exhibits the estimate frequency of firemen demand by region for the year 2013, a month of 2017, and a given day in 2016, with the three values for $\epsilon = [\ln(7), \ln(4), \ln(2)]$ (i.e., a low, a medium, and a high privacy guarantee). All three specific dates were chosen at random for illustration purposes.

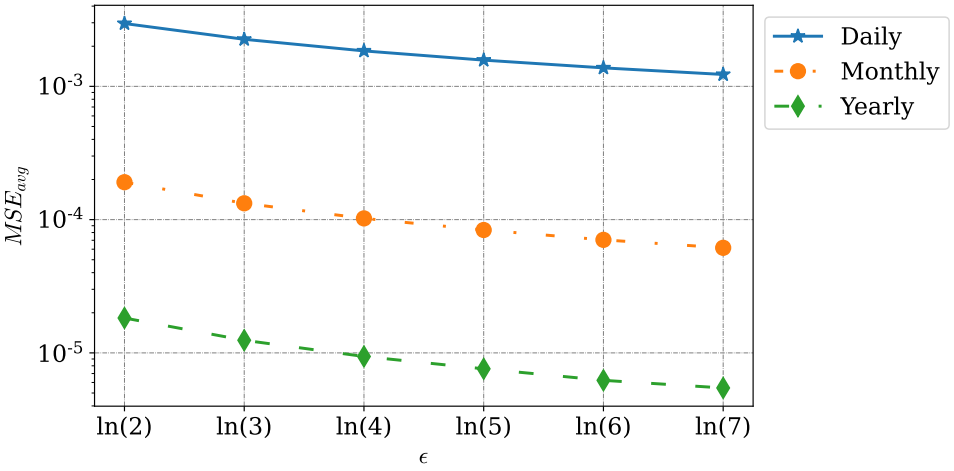

Figure 9.3: Analysis of MSE_{avg} (y-axis) varying ϵ (x-axis) for each aggregation period: daily, monthly, and yearly.

Notice that such kind of frequency estimation experiments allow evaluating the relationship between MSE_{avg} versus data size (i.e., period of analysis) according to ϵ in order to find the best privacy-utility trade-off for different applications. For instance, each scenario allows the fire brigade to have a sanitized database of intervention's region where third party companies or the human resources department itself could acquire high-utility statistics.

More specifically, as one can notice in Figures 9.3 and 9.4, estimating the frequencies of firemen demand by region with OUE can be achieved high accuracy for different aggregation periods. Intuitively, the MSE_{avg} decreases as the data size increases, with a difference of 1 order of magnitude for each aggregation scenario. In fact, this is because the variance of OUE is inversely proportional to the number of users n (cf. Eq. (2.11)). For example, for a one-year analysis, the number of interventions is at least 17333 in 2006 (and higher the other years), while the average per day is just 47 for the same year. For this reason, the utility of the data decreases for small aggregation periods.

Hence, one has to balance the application of the sanitized data. For instance, if one intends to acquire statistics per year, results are very accurate with good privacy guarantees. However, if one intends to apply machine learning tasks to this data (as presented

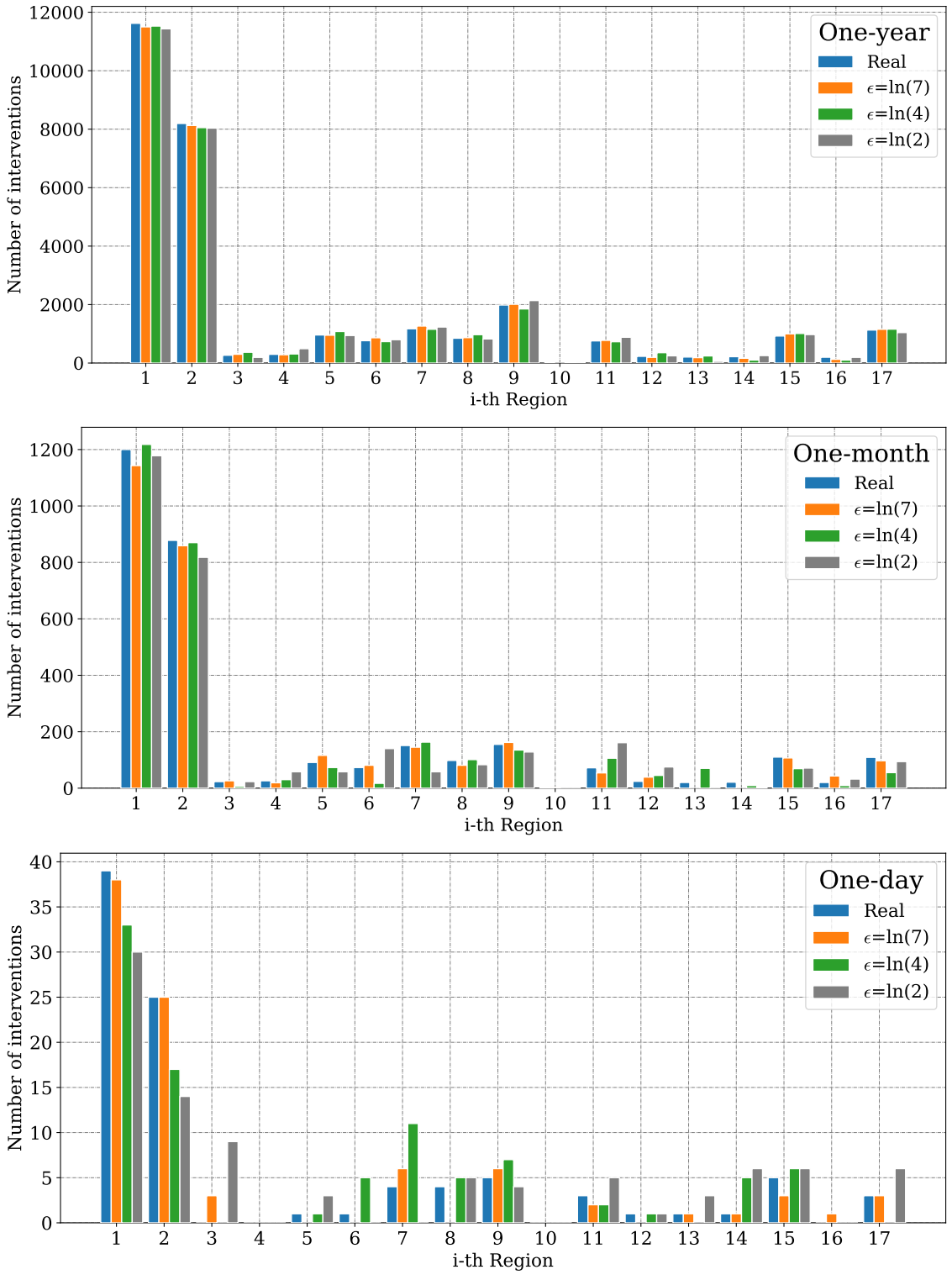

Figure 9.4: Comparison between the original and estimated firemen demand by region for one-year, one-month, and one-day periods, respectively.

in the next section), a one-day scenario is more appropriate but with a higher estimation error. For instance, in Fig. 9.3, one can see the estimated frequencies for each period,

where there are small estimation errors for the one-year scenario but considerable ones for both one-month and one-day scenarios.

Lastly, as also highlighted in the literature, the choice of ϵ depends on several factors (data size, the application domain) and one has to appropriately balance it considering the privacy of users and utility of data. In our case, as 608 cities were generalized to 17 regions, privacy could be slightly decreased (higher ϵ values) to acquire good utility for generating statistics. In the literature, common values to ϵ are within the range $0.01 - 10$ [63].

9.4/ Differentially Private Forecasting Firemen Demand by Region

The main purpose of this section is to evaluate the privacy-utility trade-off of training a state-of-the-art machine learning algorithm, namely XGBoost [76], over ϵ-DP estimated frequencies from Section 9.3, to forecast the firemen demand by 17 regions.

9.4.1/ Setup of Forecasting Experiments

Environment. All algorithms were implemented in Python 3.8.8 with XGBoost [76] and Scikit-learn [36] libraries. In all experiments, we report average results over 10 runs as LDP algorithms are randomized (i.e., the sanitized datasets).

Dataset. There are 7 datasets of frequency demand by region: the original dataset and the 6 sanitized datasets from Section 9.3, which guarantees ϵ-DP in the range $\epsilon = [\ln(2), \ln(3), ..., \ln(7)]$). More formally, each dataset $X_{(t_1,t_\tau)}$ aggregates the number of interventions per 17 regions and corresponding time period $t \in [1, \tau]$ of **daily intervals**. That is, $X_{(t_1,t_\tau)} = [\langle t_1, \mathbf{x}_1\rangle, \langle t_2, \mathbf{x}_2\rangle, ..., \langle t_\tau, \mathbf{x}_\tau\rangle]$, where $\mathbf{x}_t$ is a vector of size 17 in which each position represents the number interventions per region at time $t \in [1, \tau]$. We exclusively divided our datasets into **learning (from 2006-2017)** and **testing (the year 2018)** sets.

Temporal features. For both original and sanitized datasets, we added temporal features such as: year, month, day, weekday, year day, values (1 for 'yes', 0 for 'no') to indicate leap years, first or last day of the month, and first or last day of the year as attributes.

Forecasting methodology. In this chapter, we aim at forecasting the future firemen demand by region in the **next day**. Thus, given $X_{(t_1,t_\tau)}$, the goal is to forecast $X_{(t_{\tau+1})}$, i.e., **one-step-ahead forecasting**, which is unknown at time τ. More precisely, we only used a **single lag value**, i.e., we used the current frequency of firemen demand by region at time t as an input to predict the future frequency at time $t + 1$.

Baseline model. We established a naive forecasting technique that describes the **average** number of interventions in each day of the week per region.

Methods evaluated. In order to make a multi-forecast of firemen demand by region, the "MultiOutputRegressor" from the Scikit-learn library [36] is applied. In this regard, one regressor per target (region) is fitted using the XGBoost [76] regressor with the parameters: max_depth=3, learning_rate=0.8, and n_estimators=100. For all other parameters, we used their default values. We tuned these hyperparameters through a random search [40] optimization methodology with the following ranges per parameter: max_depth=$\{1, 2, 3, ..., 12\}$, learning_rate=[0.1,0.9], and n_estimators=$\{50, 100, 150, ..., 1000\}$. **Seven models were built: One XGBoost model trained over original data** and **six XGBoost (input perturbation-based) models trained over sanitized data** considering the aforementioned ϵ-DP range to predict the firemen demand by region for all days of 2018.

Performance metrics. All models were evaluated with standard time-series metrics, namely, RMSE and MAE, both explained in Section 3.1.5. Moreover, as it is a multi-output scenario, we only present their averaged values.

9.4.2/ FORECASTING RESULTS

Fig. 9.5 illustrates the relationship between the RMSE and MAE metrics (y-axis) with ϵ (x-axis) for the Baseline model and XGBoost ones trained over original and sanitized datasets. Lastly, Fig. 9.6 illustrates the best prediction results of a single day according to the $\epsilon = \ln(2)$-DP model. In Fig. 9.6, it is illustrated the original frequency of firemen demand by region in comparison with the predicted ones by XGBoost models trained with the original and sanitized datasets for a single day of 2018.

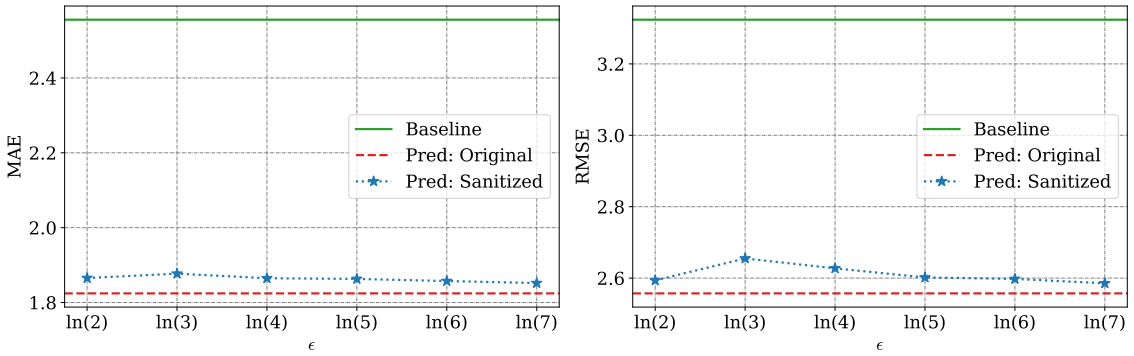

Figure 9.5: MAE and RMSE metrics for the prediction models: Baseline and XGBoost trained over original and sanitized datasets.

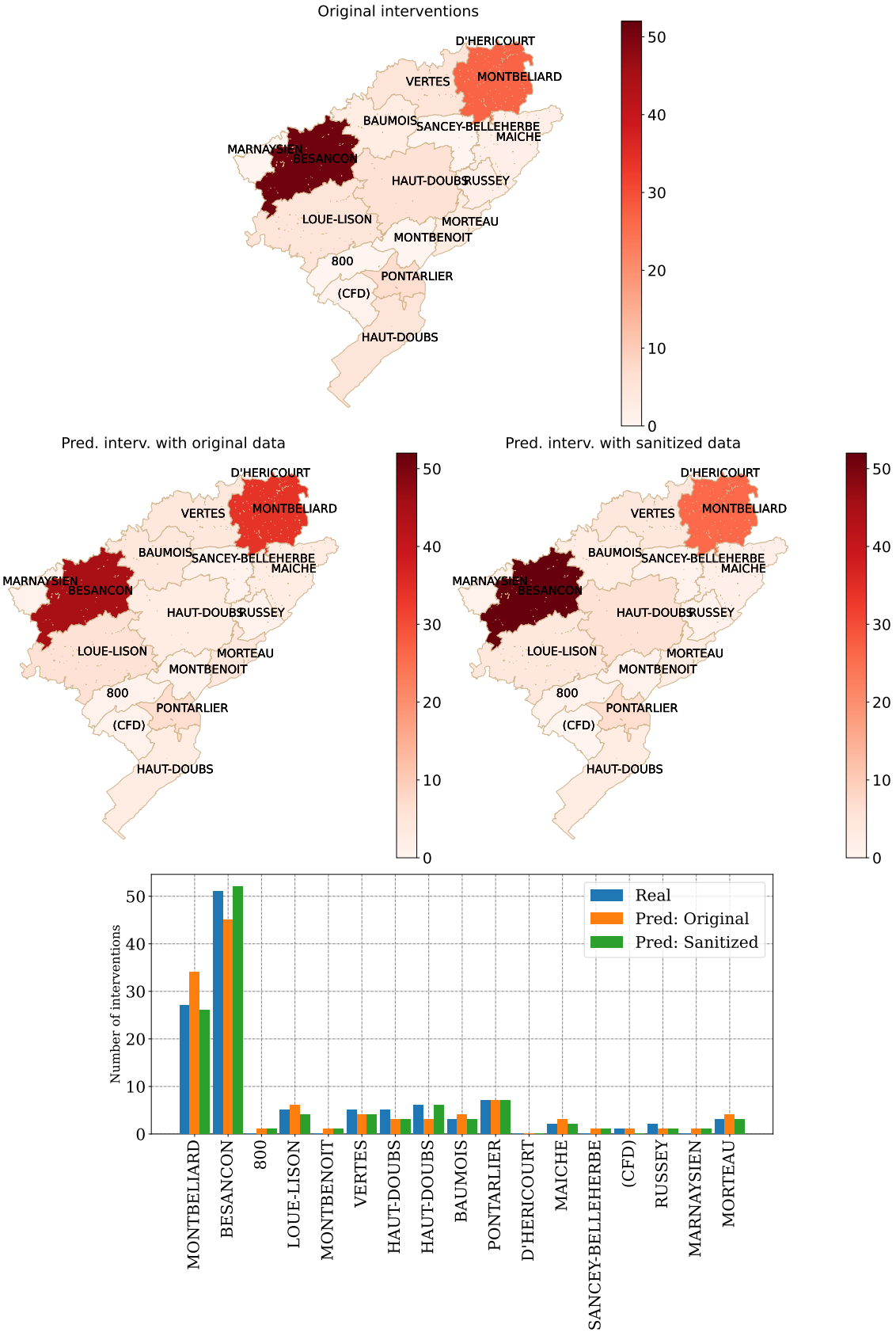

Figure 9.6: Comparison of the original and predicted firemen demand by region for a single day using XGBoost trained over original data and an $\epsilon = \ln(2)$-DP one.

First, as one can notice in Fig. 9.5, it is remarkable the improvement of the scores achieved by the XGBoost models for such complex task rather than developing a simple prediction model as the baseline (mean) assumed in this chapter. In addition, from both Figs. 9.5 and 9.6, one can notice that XGBoost models trained with sanitized data can also guarantee a good utility of the data for prediction purposes since they did not lose much utility in comparison with the model trained over original data. Indeed, this is true for the whole range of ϵ evaluated, thus, proving the effectiveness of our proposed solution, which relies on input perturbation allowing both statistical learning and forecasting tasks.

More precisely, in Fig. 9.6, it is shown for a given day of 2018 the comparison of the original and predicted firemen demand by region using the original dataset and a sanitized one with the strongest $\epsilon = \ln(2)$ tested in our experiments. As one can notice, accurate multivariate forecasts could be achieved even with a strongly sanitized dataset. With such forecasting results, the fire brigade could efficiently prepare themselves for short-, middle-, and long-term scenarios. For example, knowing that certain regions are more prospect to happen incidents, the fire brigade can better allocate the human and machinery resources as well as planning the construction of new barracks. All of these could be achieved while providing strong privacy guarantees for each intervention, using our proposed methodology.

Indeed, forecasting the operational demand is one main goal of Fire brigades (and EMS in general) to optimize their services [152, 148, 193, 172, 75, 148, 176, 162, 146, 177, 226, 154]. For instance, in [75], the authors identified that shorter ambulance response time is associated with a higher survival rate and predicted the demand of ambulances to allow their reallocation. Besides, in previous works of our research group [148, 176, 146, 177, 226, 154], several classical time-series forecasting, ML, and DL techniques have been employed to forecast the *total* number of interventions considering the whole Doubs region, in different time-slots (e.g., 1 hour, 3 hours, ...). Besides, in [148], our group also proposed to forecast the operational demand of two main regions of Doubs and by motive, by slots of 3 hours. In that work, input perturbation was considered through applying *k*-anonymity [18, 20], *l*-diversity [28], and centralized DP [27, 26, 59] algorithms to sanitize the original dataset. The main difference between the work in [148] with this chapter, is that an LDP mechanism is used to sanized row-by-row independently, which also permits the data analyst to aggregate by different slots of time (cf. Section 9.3).

9.5/ CONCLUSION

In this chapter, we proposed a privacy-preserving methodology based on generalization and LDP, which would allow EMS to use and/or share the sanitized database for both **sta-**

tistical learning and **forecasting tasks** on the frequency of firemen demand by region. Indeed, while there are examples of EMS open data publication [45, 111], we believe that more attention should be given to the privacy of the victims concerned. For instance, EMS should not blindly generalized the number of interventions per month/week and city, as there could be many cases of unique interventions (e.g., see Tables 9.2 and 9.1). Moreover, in a more non-private case, we argue that publishing the **precise information about the time, the location, and the reason of the emergency** (e.g., as in [8]), **could increase the possibility of breaching someone's privacy**.

Therefore, in our solution, we propose that both the time and the location be generalized, the former by *day* and the latter by big regions (e.g., we generalize 608 cities to 17 regions in Fig. 9.2). In addition to generalization, we also propose that an ϵ-LDP protocol (cf. Section 2.4) be applied to each interventions' region (i.e., row-by-row sanitization), thus, enhancing the privacy of users. In this chapter, we used the state-of-the-art OUE [108] protocol for single attribute frequency estimation. As shown in the results of Section 9.3, the OUE mechanism can adequately estimate the firemen demand by region with a good level of privacy guarantees for all three aggregation periods (see Fig. 9.4).

Moreover, as shown in Section 9.4, it is possible to forecast the future firemen demand by region with sanitized data as well as with the original data (cf. Fig 9.5). More specifically, the work in this chapter shows that EMS data, which is sensitive but can be very useful, can be properly sanitized to avoid information leakage while remaining useful for both statistical learning (cf. Fig. 9.4) and forecasting (cf. Fig. 9.6) purposes. Lastly, while this chapter focused on aggregate information, thus, applying generalization and LDP protocols for frequency estimation, the next Chapter 10 investigates how to sanitize the coordinates (i.e., latitude and longitude) of the emergency's location, focusing on a different problem, i.e., predicting the response time of each ambulance.

10

Preserving Emergency's Location Privacy to Predict Response Time

In Chapters 1, we have reviewed our second motivating project concerning the SDIS 25 (i.e., an EMS in France) and in Chapter 9, we have proposed an LDP-based methodology focusing on both *statistical learning* and *machine learning forecasting* on the frequency of firemen demand by region. In this Chapter 10 and in Chapter 11, **following our collaboration with Selene Cerna**, we focus our attention on a different problem, which concerns the **response time of EMS to each emergency**. Indeed, many victims require care within adequate time (e.g., cardiac arrest) and, thus, improving response time is vital. In this context, the **location** of the emergency is a determinant factor of EMS response time since it defines, e.g., the distance between the EMS center and the emergency scene.

With these elements in mind, we asked ourselves, is the **precise location really necessary to be used as a predictor of an ML model that predicts ambulance response times?** In fact, we still consider that EMS intend to share a sanitized version of their data, such that third parties could build decision-support tools to optimize the EMS service. So, **in collaboration with Selene Cerna**, we propose in this chapter to use the geo-indistinguishability [46] LDP model to sanitize each emergency scene independently (i.e., row-by-row). In addition, there are many other predictors that may also be 'perturbed', e.g., the calculated distance between both the EMS center and the emergency scene; the estimated travel time, the city, and so on. Thus, thanks to the post-processing properties of DP [59], EMS could use and/or share the sanitized data with third parties to develop ML-based decision-support tools. We invite the reader to refer to Chapter 2 for the background on LDP and geo-indistinguishability. Lastly, similar to Chapter 5 and 9, we highlight that although we apply an *LDP-based* mechanism, this is a centralizer data owner (i.e., EMS) that applies the geo-indistinguishability protocol on its servers, thus, providing centralized privacy guarantees for users.

10.1/ INTRODUCTION

Ambulance response time (ART) is a key component for evaluating pre-hospital EMS operations. ART refers to the period between the EMS notification and the moment an ambulance arrives at the emergency scene [121, 144]. In many urgent situations (e.g., cardiovascular emergencies, trauma, or respiratory distress), the victims need first-aid treatment within adequate time to increase survival rate [41, 144, 188, 121, 75, 157] and, hence, improving ART is vital.

One important factor of ART is the *location* of the intervention [82, 41, 158, 144, 33], e.g., in dense urban areas, the distance may be short, but the travel time may be longer due to traffic congestion. On the other hand, travel distance and travel time may be longer for rural areas. In other words, the location information is of great importance for the prediction of travel time and, naturally, ART [82, 33]. As also mentioned in Chapter 9, **the location of an emergency**, on the other hand, **is considered sensitive information** since it might identify who received assistance and for what purpose. For example, attackers with auxiliary information may correctly deduce that a weakened person activated the EMS if they know that one intervention took place in front of their residence. The attackers may then exploit this knowledge for their own benefit.

In this chapter, we propose to sanitize, independently, each emergency location data with geo-indistinguishability (GI) [46] (cf. Section 2.5), which is based on the state-of-the-art DP [27, 26, 59] model. Indeed, we aim to evaluate the effectiveness of several values of ϵ (i.e., the privacy budget) to sanitize emergency location data with GI and train ML-based models to predict ART. In other words, this is a practical evaluation of GI on a real-world EMS task. This way, EMS would only use and/or share sanitized data with third parties to *train* and develop ML-based decision support systems, thus, protecting their victims if there are data leakages [228] or if the built ML model is subject to membership inference attacks and data reconstruction attacks [105, 104, 145].

In our context, besides the own location, with the exact coordinates of both SDIS 25 centers and the emergency scenes, one can retrieve important features such as the distance and estimated travel time. However, if the location is sanitized via GI, many other explanatory variables (e.g., distance, travel time, city) would be 'perturbed' too. As reviewed in Section 3.2.1, training ML models with sanitized data is also known as *input perturbation* [32, 31].

We perform our experiments on the SDIS 25 preprocessed dataset named **ART-DB** from Section 3.3.4. The ART-DB contains information about 186130 **dispatched ambulances** from SDIS 25 centers that attended 182700 EMS interventions from 2006 up to June 2020. To the author's knowledge, this is the first work to assess the impact of geo-indistinguishability on sanitizing the location of emergency scenes when training the ML

model for such an important task.

The remainder of this chapter is organized as follows. In Section 10.2, we describe the sanitization of emergency scenes with GI and the experimental setup. In Section 10.3, we present the results of our experiments and its discussion including related work. Lastly, in Section 10.4, we present the concluding remarks. The development, results and discussion presented in this chapter were published in a full article [212] in the Mathematical and Computational Applications journal.

10.2/ Materials and Methods

In this section, we present the GI-based sanitization of emergency location data (Section 10.2.1) and the experimental setup (Section 10.2.2).

10.2.1/ Preserving emergency location privacy with geo-indistinguishability

To preserve geo-indistinguishability of each emergency scene, we apply the polar Laplace mechanism in Alg. 2 presented in Chapter 2 to the *Location* attribute of **each intervention**. The codes we developed and used for all experiments are available in a Github repository[1]. More specifically, even if the ART-DB is per ambulance dispatch (i.e., 186130 ambulances), we used the same sanitized value per intervention (i.e., 182700 unique interventions). Although in [46] the authors propose two further steps to Alg. 2, i.e., discretization and truncation, both steps can be neglected in our context. This is, first, because SDIS 25 may also help other EMS outside the Doubs region, and second, we assume that any location in the continuous plane can be an emergency scene. While reporting an approximate location in the middle of a river may not have much sense in location-based services, in an emergency dataset with approximate locations, this may indicate an urgency for someone who drowned in the river, for example.

We used five different levels for the privacy budget $\epsilon = l/r$, where l is the privacy level we want within a radius r. Table 10.1 exhibits the five different levels of privacy, selected similar to the original GI paper [46]. For the sake of illustration, Figure 10.1 exhibits three maps of the Doubs region with the points of original location (left-hand plot), $\epsilon = 0.005493$-GI location (middle plot), and $\epsilon = 0.002747$-GI location (right-hand plot). As one can notice, with an intermediate privacy level ($l = \ln(3), r = 400$), locations are more spread throughout the map while with a lower privacy level ($l = \ln(3), r = 200$), locations approximate the real clusters.

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

$\epsilon = l/r$	l	r (meters)
0.005493	$\ln(3)$	200
0.002747	$\ln(3)$	400
0.001155	$\ln(2)$	600
0.000866	$\ln(2)$	800
0.000693	$\ln(2)$	1,000

Table 10.1: Values of $\epsilon = l/r$ for sanitizing emergency location data with GI.

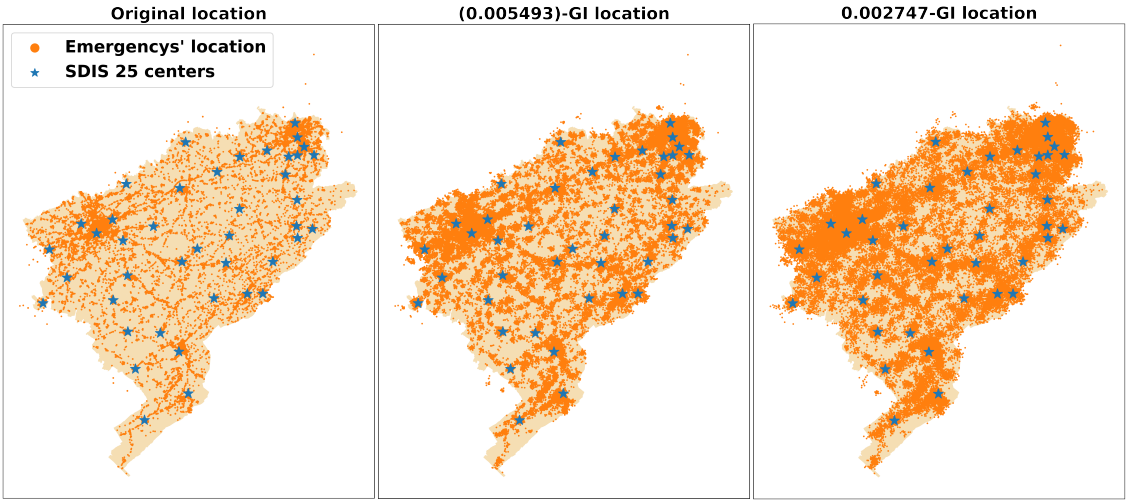

Figure 10.1: Emergency locations and SDIS 25 centers throughout the Doubs region: original data (left-hand plot), $\epsilon = 0.005493$-GI data (middle plot), and $\epsilon = 0.002747$-GI data (right-hand plot).

Moreover, with the new *Location* values of each intervention, we also reassigned the city, the district, and the zone **when applicable**. In addition, we recalculated the following features associated with it: the great-circle distance [3], the estimated driving distance, and estimated travel time. The latter two features were recalculated with the open source routing machine (OSRM) API [35], which only considers roads, i.e., if the obfuscated location is in the middle of a farm, the closest route estimates the driving distance and travel time until the closest road. We also highlight that if the new coordinates of the emergency scene indicate a location closer to another SDIS 25 center, even in real life, it would not imply that this center took charge of the intervention. Therefore, the *center* attribute **was not** 'perturbed'.

To show the impact of the noise added to the *Location* attribute, Table 10.2 exhibits the percentage of time that categorical attributes (zone, city, and district) were 'perturbed' (i.e., reassigned); the mean and std values of the great-circle distance attribute (considering the SDIS 25 center and the emergency scene) and its Pearson correlation coefficient [7] with the ART variable (Corr. ART). In Table 10.2, we report the mean(std)

values since we repeated our experiments with 10 different seeds (i.e., DP algorithms are randomized). Although we did not include the estimated driving distance and estimated travel time from OSRM API in this analysis, in preliminary tests, we noticed that these two features follow a similar pattern as the great-circle distance attribute.

From Table 10.2, one can notice that many features are perturbed due to sanitization of emergency's location with GI. With high levels of ϵ (i.e., less private), the city and the zone suffer low 'perturbation'. On the other hand, district is reassigned many times as it is geographically smaller than the others. For example, in the fourth row Table 10.2, when $\epsilon = 0.000866$, the city is already reassigned more than 50% of the time and the district about 74% of the time. Moreover, one can notice that the mean and std values of the great-circle distance increase as the ϵ parameter decreases (i.e., more private). Because $\epsilon = l/r$, making l smaller and/or r higher, the stricter ϵ becomes, and therefore more noise is added to the original locations. Besides, the Pearson correlation coefficient between the great-circle distance with the ART variable decreases as ϵ becomes smaller.

Data	**Zone**	**City**	**District**	**Great-circle Dist. (km)**		
	'perturbation' (%)			Mean	std	Corr. ART
Original	-	-	-	3.44	3.72	0.369
$\epsilon = 0.005493$	5.20(0.05)	7.68(0.06)	25.8(0.05)	3.48(1e-3)	3.72(7e-4)	0.367(2e-4)
$\epsilon = 0.002747$	11.3(0.05)	17.6(0.10)	41.5(0.12)	3.57(1e-3)	3.72(1e-3)	0.362(2e-4)
$\epsilon = 0.001155$	28.1(0.06)	42.3(0.10)	66.2(0.09)	4.03(3e-3)	3.74(3e-3)	0.335(5e-4)
$\epsilon = 0.000866$	35.5(0.10)	52.4(0.11)	74.0(0.11)	4.38(3e-3)	3.81(4e-3)	0.313(1e-3)
$\epsilon = 0.000693$	41.4(0.12)	60.3(0.09)	79.4(0.05)	4.77(6e-3)	3.92(5e-3)	0.288(1e-3)

Table 10.2: Percentage of perturbation for categorical attributes (city, zone, and district) according to ϵ and statistical properties (mean and std values and correlation with ART) of the original and GI-based datasets for the great-circle distance attribute. Mean(std) values are reported since we repeated our experiments with 10 different seeds.

10.2.2/ SETUP OF EXPERIMENTS

Four state-of-the-art ML techniques have been used in our experiments, **to predict the scalar ART outcome** in a regression framework. More precisely, we compared the performance of two state-of-the-art ML techniques based on decision trees, which are known for their high performance (and speed) with tabular data (i.e., LGBM [76] and XGBoost [100]); a traditional and well-known deep learning (i.e., MLP [79, 69]), and a classical statistical method that can perform both variable selection and regularization (i.e., LASSO [15]). All these methods have been revised in Section 3.1.3

Because in Table 10.2 there are low variations (i.e., small std values) on all features that depend on the sanitized location, we ran our experimental validation only once. As detailed in Section 3.3.4, in our experiments, **each sample corresponds to one ambu-**

lance dispatch, in which there are **temporal features** (e.g., hour, day), , **traffic data** (i.e., indicators from [244]), hourly **weather data** (e.g., temperature, pressure, ..., from [246]), **location-based features** (latitude, longitude, district, city, and the zone), and **computed features** (e.g., the distance between the SDIS 25 center and the emergency scene, estimated travel time, estimated driving distance, where the two latter are from [35]). **The scalar target variable is the ART in minutes, which is the time measured from the SDIS 25 notification to the ambulance's arrival on-scene**.

In addition, the ART-DB was preprocessed by Selene Cerna as follows. All numerical features (e.g., temperature) were standardized using the *StandardScaler* function from the Scikit-learn library [36]. Categorical features (e.g., center, zone, hour) were encoded using mean encoding, i.e., the mean value of the ART variable with respect to each feature (considering the training set only). **The target variable, namely ART,** was kept in its original format (minutes) since no remarkable improvement was achieved with scaling.

With these elements in mind, we divided the ART-DB into **training (years 2006-2019)** and **testing (six months of 2020)** sets to evaluate our models. Thus, five models per ML technique (i.e., XGBoost, LGBM, MLP, and LASSO) were built to predict ART on each month of 2020 using the sanitized (training) datasets with different levels of ϵ-GI location data (*cf.* Table 10.1). All models were trained continuously, i.e., at the end of each month, the new known data were added to the training set after sanitization with ϵ-GI. **Lastly, all models were tested with original data. On the one hand, this would prevent having in real-life a sanitized location that would compromise the EMS response time. On the other hand, each time the model is re-fitted (or retrained), the new known data should also be sanitized with ϵ-GI.** In addition, for comparison purposes, we also trained and evaluated one additional model per ML technique with original data. In this chapter, the models were evaluated using the following regression metrics: RMSE, MAE, MAPE, and R^2, all presented in Section 3.1.5.

Results for each metric were calculated using data from the 6 months evaluation period. The RMSE metric was also used during the hyperparameters tuning process via Bayesian optimization (BO), explained in Section 3.1.6. To this end, we used the HYPEROPT library [47] with 100 iterations for each model. Table 10.3 displays the range of each hyperparameter used in the BO, as well as the final configuration used to train and test the models.

10.3/ RESULTS AND DISCUSSION

In this section, we present the results of our experimental validation (Section 10.3.1) and a general discussion (Section 10.3.2) including related work and limitations.

Model	Search space	Final configuration per dataset					
		Original	$\epsilon = 0.005493$	$\epsilon = 0.002747$	$\epsilon = 0.001155$	$\epsilon = 0.000866$	$\epsilon = 0.000693$
XGBoost	max_depth: [1, 10]	9	9	6	6	9	9
	n_estimators: [50, 500]	465	465	130	235	465	465
	learning_rate: [0.001, 0.5]	0.0265	0.0265	0.0858	0.0486	0.0265	0.0265
	min_child_weight: [1, 10]	5	5	7	7	5	5
	max_delta_step: [1, 11]	4	4	3	4	4	4
	gamma: [0.5, 5]	3	3	0	2	3	3
	subsample: [0.5, 1]	0.8	0.8	1	1	0.8	0.8
	colsample_bytree: [0.5, 1]	0.5	0.5	0.5	0.5	0.5	0.5
	alpha: [0, 5]	2	2	1	2	2	2
LGBM	max_depth: [1, 10]	7	8	10	8	8	6
	n_estimators: [50, 500]	355	326	477	250	80	441
	learning_rate: [1e-4, 0.5]	0.0188	0.0098	0.0164	0.0285	0.0586	0.0300
	subsample: [0.5, 1]	0.54066	0.5228	0.6138	0.6699	0.6732	0.5812
	colsample_bytree: [0.5, 1]	0.5160	0.5575	0.5204	0.6870	0.5507	0.5451
	num_leaves: [31, 400]	400	192	245	398	132	95
	reg_alpha: [0, 5]	4	0	5	0	1	4
MLP	Dense layers: [1, 7]	7	3	4	6	6	6
	Number of neurons: $[2^8, 2^{13}]$	2^{10}	2^{12}	2^{12}	2^9	2^{12}	2^9
	Batch size: [32, 168]	140	80	48	82	70	44
	Learning rate: [1e-5, 0.01]	0.00265	0.00124	0.0099	0.0099	0.0094	0.0077
	Optimizer: Adam	Adam	Adam	Adam	Adam	Adam	Adam
	Epochs: 100	100	100	100	100	100	100
	Early stopping: 10	10	10	10	10	10	10
LASSO	alpha: [0.01, 2]	0.0205	0.0307	0.0105	0.0100	0.0112	0.0107

Table 10.3: Search space for hyperparameters by ML model and the final configuration obtained for predicting ARTs per dataset.

10.3.1/ PRIVACY-PRESERVING ART PREDICTION

Figure 10.2 illustrates the impact of the level of GI for each ML model to predict ART according to each metric. As one can notice in this figure, for XGBoost, LGBM, and LASSO, there were minor differences between training models with original location data or sanitized ones. On the other hand, models trained with MLP performed poorly with GI-based data. In addition, by analyzing models trained with original data, while the smaller RMSE for LASSO is about 5.65, for more complex ML-based models, RMSE is less than 5.6, achieving 5.54 with XGBoost and LGBM. In comparison with the results of existing literature, lower R^2 scores and similar RMSE and MAE results were achieved in [158] to predict ART while using original location data only.

Indeed, among the four tested models, LGBM and XGBoost achieve similar metric results while favoring the LGBM model. Thus, Figure 10.3 illustrates the BO iterative process for LGBM models trained with original and sanitized data according to the RMSE metric (left-hand plot); and ART prediction results for 50 dispatched ambulances in 2020 out of 8,709

ones (right-hand plot) with an LGBM model trained with original data (Pred: original) and with two LGBM models trained sanitized data, i.e., with $\epsilon = 0.005493$ (low privacy level) and with $\epsilon = 0.000693$ (high privacy level).

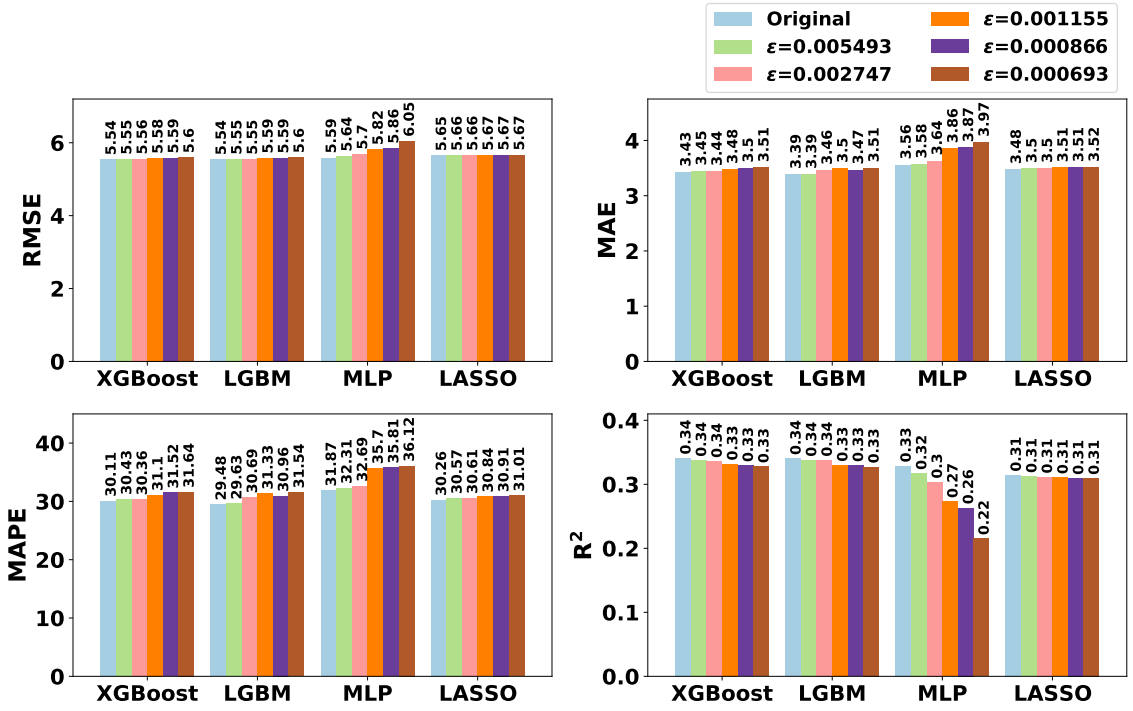

Figure 10.2: Impact of the level of ϵ-geo-indistinguishability for each ML model to predict ART according to each metric.

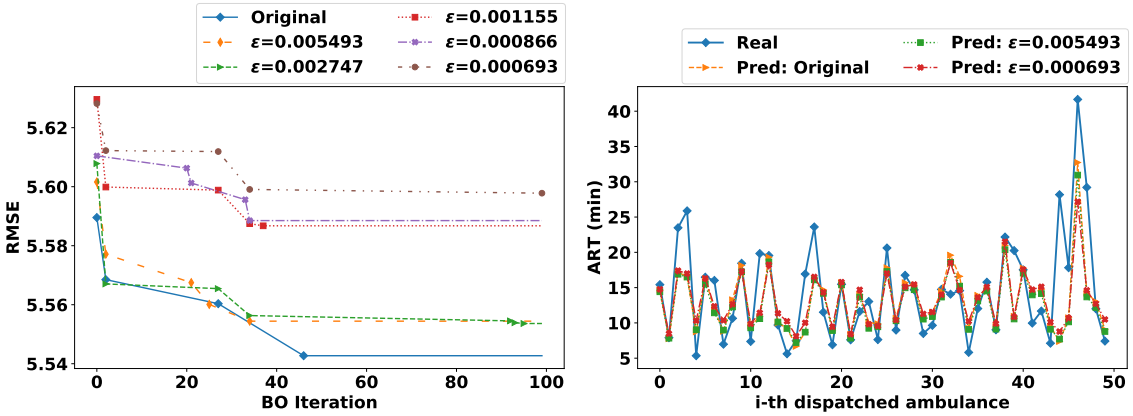

Figure 10.3: The left-hand plot illustrates the hyperparameters tuning process via Bayesian optimization with 100 iterations for LGBM models trained with original data and sanitized ones. The right-hand plot illustrates the prediction of ARTs with LGBM models trained with original data and with sanitized ones.

As one can notice in the left-hand plot of Figure 10.3, once data are sanitized with different levels of ϵ-GI, the hyperparameters optimization via BO is also perturbed. This way, local minimums were achieved in different steps of the BO (i.e., the last marker per curve

indicates the local minimum). For instance, even though $\epsilon = 0.002747$ is more strict than $\epsilon = 0.005493$, results were still better for the former since, in the last steps of BO, three better local minimums were found. Besides, prospective predictions were achieved with either original or sanitized data. For instance, in the right-hand plot of Figure 10.3, even for the high peak-value of ART around 40 minutes, LGBM's prediction achieved some reasonable estimation. Although several features were perturbed due to the sanitization of the emergency scene (e.g., city, zone, etc), the models could still achieve similar predictions as the model trained with original location data.

10.3.2/ Discussion and Related Work

The medical literature has mainly focused attention on the analysis of ART [41, 21, 30] and its association with trauma [19, 144] and cardiac arrest [121, 157, 188], for example. To reduce ART, some works propose reallocation of ambulances [175, 75], operation demand forecasting [152, 148, 193, 172, 75, 148, 176, 146, 177, 226, 154], travel time prediction [33], simulation models [39, 23], and EMS response time predictions [33, 158]. The work in [158] propose a real-time system for predicting ARTs for the San Francisco fire department, which closely relates to our work in this chapter. The authors processed about 4.5 million EMS calls utilizing original location data to predict ART using four ML models, namely linear regression, linear regression with elastic net regularization, decision tree regression, and random forest. However, no privacy-preserving experiment was performed because the main objective of their paper was proposing a scalable, ML-based, and real-time system for predicting ART. Besides, we also included weather data that the authors in [158] did not consider in their system, which could help to recognize high ARTs due to bad weather conditions, for example.

Because most of EMS data are personal and confidential (e.g., location, reason), there is a need for privacy-preserving techniques for processing and using these data. In this chapter, even if the intervention's *reason* could be an indicator of the call urgency, we did not consider this sensitive attribute in our data analysis nor privacy-preserving prediction models. This is because, for SDIS 25, the ARTs limits are defined by the zone [178]. Additionally, we also did not include the victims' personal data (e.g., gender, age) in our predictions or analysis since, during the calls, the operator may not acquire such information, e.g., when a third party activates the SDIS 25 for unidentified victims. This way, we focused our attention on the **location privacy of each intervention**.

Indeed, location privacy is an emerging and active research topic in the literature [163, 77, 37, 91, 46, 240] as publicly exposing users' location raises major privacy issues. To address this location privacy issue, in this chapter, we sanitized each emergency location using the state-of-the-art GI [46] model. As highlighted in [46], attackers in LSBs may

have side information about the user's reported location, e.g., knowing that the user is probably visiting the Eiffel Tower instead of swimming in the Seine river. However, this does not apply in our context because someone may have drowned and EMS had to intervene. Similarly, even for the dataset with intermediate (and high) privacy in which locations are spread out in the Doubs region (*cf.* map with 0.005493-GI location in Figure 10.1), someone may have been lost in the forest and EMS would have to interfere. For these reasons, using (or sharing datasets with) approximate emergency locations (e.g., sanitized with GI) is a prospective direction since many locations are possible emergency scenes. Indeed, we are not interested in hiding the emergency's location completely since some approximate information is required in order to retrieve other features (e.g., city, zone, estimated distance) to use for predicting ART.

With the differentially private input perturbation setting adopted in this chapter, data are protected from data leakage and are more difficult to reconstruct, for example. For instance, the authors in [191, 98] investigate how input perturbation through applying controlled Gaussian noise on data samples can guarantee (ϵ, δ)-DP on the final ML model. This means, since ML models are trained with perturbed data, there is a perturbation on the gradient and on the final parameters of the model too.

In this chapter, rather than Gaussian noise, the emergency scenes were sanitized with Alg. 2 explained in Chapter 2, i.e., adding two-dimensional Laplacian noise centered at the exact user location $x \in \mathbb{R}^2$. In addition, this sanitization also perturb other associated and calculated features such as: city, district, zone (e.g., urban or not), great-circle distance, estimated driving distance, and estimated travel time (*cf.* Table 10.2). As well as the optimization of hyperparameters, i.e., once data are differentially private, one can apply any function on it and, therefore, we also noticed perturbation on the BO procedure. Yet, as shown in the results, prospective ART predictions were achieved with either original or sanitized data. What is more, even with a high level of sanitization ($\epsilon = 0.000693$) there was an adequate privacy-utility trade-off. According to [14], if the mean absolute percentage error (i.e., MAPE) is greater than 20% and less than 50%, the forecast is reasonable, which is the results we have in this chapter with MAPE around 30%.

10.4/ CONCLUSION

In this chapter, we aimed to predict the response time that each center equipped with ambulances had to an event, which could be used as an intelligent decision-support system to dynamically select the center to deploy ambulances. However, we also took into consideration that the emergency locations are sensitive data, requiring proper sanitization. Therefore, this interdisciplinary work aimed to evaluate the effectiveness of predicting ARTs with ML models trained over sanitized location data with different levels of ϵ-geo-

indistinguishability.

As shown in the results, the sanitization of location data and the perturbation of its associated features (e.g., city, distance) had certain impact on data utility (see Table 10.2) but no considerable effect on predicting ART (see Fig. 10.3). With these findings, EMS may prefer using and/or sharing sanitized datasets to avoid possible data leakages, membership inference attacks, or data reconstruction attacks, for example [228, 105, 104, 145]. **In conclusion, while predicting ART might allow EMS to save more lives, we notice that it is also possible to do so while preserving the victims' location privacy.**

Lastly, on the one hand, this chapter focused on **response time** taking into consideration the **recommended times** SDIS 25 ambulances should arrive on-scene (e.g., for Z1 the ART should be ≤ 10 minutes) [178]. The next Chapter 11 proposes a privacy-preserving solution to **response time** taking into consideration the urgency level of the intervention through predicting if each victim will die.

11

PRIVACY-PRESERVING PREDICTION OF VICTIM'S MORTALITY

In Chapter 10, we have started to focus on EMS **response time** to emergencies. In this last chapter of contribution, we continue in this direction from another perspective: **Although SDIS 25 ARTs depend mainly on the zone [178], is there a way to recognize or of being aware that an emergency will require *priority attention*?** To answer this question, **with Selene Cerna**, we proposed a methodology based on ML techniques to predict the victims' mortality **using data gathered from the start of the emergency call until the SDIS 25 is notified**. Within this interval of interest, there are data about the call processing times, operators' and victims' personal data; the location of the emergency, and so on. In other words, there are two entities we will be concerned with, namely, **call center operators** and **victims** regarding **privacy**. Similar to Chapters 9 and 10, we still consider that EMS intend to share an anonymized version of their data, such that third parties could build decision-support tools to optimize the EMS service. However, **differently of a single sensitive attribute** (i.e., *only location* in Chapters 9 and 10), **there are several personal attributes concerning victims and operators**. Therefore, **in this chapter**, we evaluated the **privacy-utility trade-off** of ML models trained over anonymized data using either the k-anonymity model (cf. Section 2.2) or of a differentially private algorithm [120] that produces truthful data output. **Throughout this chapter, we slightly abuse of our notation and use the terms *anonymized/anonymization* for both *k-anonymity* and DP (instead of anonymized/sanitization) guarantees** (cf. Section 2.1). We invite the reader to refer to Chapter 2 for the background on both k-anonymity [18, 20] and DP [27, 26, 59] models.

11.1/ INTRODUCTION

As reviewed in Chapters 1, 3, 9, and 10, EMS are a key component of healthcare systems around the world. An important measurement of their quality is their response time,

which is measured from the time the EMS is notified to the time an ambulance arrives at the emergency scene (cf. Chapter 10). In fact, shorter ambulance response times are potential contributors to higher survival rates [188, 144, 121, 75, 19, 157] since every second is a matter of life. For instance, **the response time also depends on how and by whom the call is processed** in the EMS center [84]. For this reason, there is a need to optimize these services and take advantage of plenty of data gathered throughout the years in hospitals and EMS.

In this chapter, we consider as *interval of interest* the period comprising the time where the SDIS 25 call center's phone starts to ring until some center(s) is notified to handle the intervention or the call ends. With all accessible data within this interval (e.g., victims and operators data, call processing times, ...), **the purpose of this chapter is to evaluate the privacy-utility trade-off of training ML models over anonymized data to predict the victims' mortality**. Therefore, there are two entities we are concerned with, namely, *call center operators* and *victims* with regard to privacy.

For instance, with the raw dataset containing direct identifiers (e.g., names), one straightforward question as: "*Is there any operator linked with an increased ratio of victims' death*?" can be easily computed, which compromises the operators' privacy and can lead to social and/or economical damages. Similarly, one can easily access the reason for the intervention (e.g., cardiac arrest) and use this information to jeopardize the victims' privacy through discrimination in health insurance, for example. Besides, as reviewed in Section 2.2, even by excluding direct identifiers, both victims' and operators' identities are still at risk of being retrieved [72]. Indeed, attributes such as gender, age, and ZIP code (a.k.a. quasi-identifiers – QIDs) can be combined with public data to reidentify individuals [18, 20].

For instance, on analyzing the Vic_Mort-DB from Section 3.3.5, considering *victims*, by combining three available QIDs (gender, age, and city), one can find about 22000 cases with the trivial $k = 1$-anonymity level [18, 20]. This means, in some cities with low population density, it would not be difficult to find out the person who needed help by knowing their gender and age. Similarly, combining four QIDs considering *operators* (gender, age, grade/career, and seniority) leads to a similar output with many unique rows. One exception is that there is a set of operators, and each row represents an *event* of who treated the emergency call. This reinforces the need for applying privacy-preserving techniques to protect the users' privacy.

Therefore, in this chapter, we assessed the effectiveness of anonymizing the Vic_Mort-DB from Section 3.3.5 with two state-of-the-art privacy techniques, namely, *k*-anonymity [18, 20] and DP [27, 26, 59] before training any ML model to predict the victims' mortality. The Vic_Mort-DB has information about 177883 **victims** that the SDIS 25 attended from January 2015 to December 2020. To the author's knowledge, this is the first work to

assess the impact of privacy-preserving techniques on predicting the victims' mortality. Indeed, while these predictions may allow SDIS 25 (or EMS in general) to save more lives, we notice that it is also possible to do so with anonymized datasets, which preserves both victims' and operators' privacy.

The remainder of this chapter is organized as follows. In Section 11.2, we present the experimental setup, our results, discussion, and we review related work. Lastly, in Section 11.3, we present the concluding remarks. The proposed privacy-preserving methodology to predict the victims' mortality (and their transportation to health facilities not approached here) developed with Selene Cerna, part of the results/discussion of Section 11.2 were accepted as a full article [221] in the Transactions on Industrial Informatics journal.

11.2/ EXPERIMENTAL VALIDATION

We divide this section in the following way. First, we describe general settings for our experiments (Section 11.2.1). Next, we present the development and evaluation of privacy-preserving ML models (Section 11.2.2). Lastly, we discuss our work and review related work (Section 11.2.3).

11.2.1/ GENERAL SETUP OF EXPERIMENTS

Environment. All algorithms were implemented in Python 3.8.8 with XGBoost [76] and Scikit-learn [36] libraries. The anonymization methods were implemented with the ARX[1] tool [70].

ML model evaluated. With Selene Cerna, two ML models and two DL models have been compared with the original data, i.e., with Vic_Mort-DB. The most performing method was XGBoost. Therefore, **only XGBoost** will be used in this chapter to evaluate its privacy-utility trade-off of being trained over anonymized data.

Dataset. We utilize the Vic_Mort-DB from Section 3.3.5 divided into exclusively **learning** (from 2015-2019 with $n_l = 149321$ victims) and **testing** (the year 2020 with $n_t = 28562$ victims) sets.

Privacy models evaluated. We compared the effectiveness of both k-anonymity [18, 20] and DP [27, 26, 59] models, both presented in Chapter 2. The differentially private model of ARX was proposed in [120], namely, *SafePub*, which produces truthful data output. More precisely, DP is ensured by sampling, in which the sampling probability depends on ϵ, and data are released in a generalized form that also satisfies k-anonymity (where k

[1] https://arx.deidentifier.org/

depends on ϵ and δ). Also, we highlight that *both privacy models were applied only in the learning set* and, hence, the *testing set was transformed* using the final generalization hierarchies.

In our experiments, we vary the ϵ parameter in the range $\epsilon = [0.2, 0.4, 0.6, 0.8, 1.0]$ and we fix $\delta = 10^{-6} \ll 1/n_l$. *With these parameters, the differentially private learning sets* were sub-sampled from $n_l = 149321$ to $n_l^* = [24621, 45038, 61841, 76418, 88663]$ samples and, *besides DP guarantees*, *k*-anonymity is also satisfied with $\mathbf{k} = [62, 62, 65, 70, 74]$, respectively. Thus, *for a fair comparison between the two privacy models*, we also set $\mathbf{k} = [62, 62, 65, 70, 74]$ when applying the *k*-anonymity model.

Generalization approach. The following (generalization) transformations were considered to anonymize each information concerning the victim (Vic.) and operator (Ope.):

- *Age (Vic. and Ope.)* by intervals of growing amplitude: 10, 20, 40, 80, total suppression (*);
- *Gender (Vic. and Ope.)* by total suppression (*);
- *Vic. City ID* by masking (five) digits: 2222*, 222**, 22***, 2****, total suppression (*);
- *Ope. Seniority (in days)* by intervals of growing amplitude (about 6 months): 180, 360, 720, ..., total suppression (*);
- *Ope. Grade* by total suppression (*).

Experimental evaluation. Eleven models were built. **One** XGBoost model trained over original data, **five** XGBoost (input perturbation-based) models trained over DP data considering the aforementioned ϵ-DP range, and **five** XGBoost (input perturbation-based) models trained over *k*-anonymous data considering the aforementioned **k** range. To optimize XGBoost hyperparameters, we applied Bayesian optimization [47] (explained in Section 3.1.6) with 100 iterations, with the following specification: n_estimators [50-1000], learning_rate [0.001-0.5], max_depth [1-20], colsample_by_tree [0.2-1], and scale_pos_weight [20-60]. For all other parameters, we used their default values.

Performance metrics. All XGBoost models were evaluated with standard binary classification metrics, namely, ACC (accuracy) and MF1 (macro f1-score), both explained in Section 3.1.5. The MF1 metric was also used as the objective function for the Bayesian optimization.

11.2.2/ PRIVACY-PRESERVING BINARY CLASSIFICATION OF VICTIMS' MORTALITY

Fig. 11.1 illustrates the relationship between the MF1 and ACC metrics (y-axis) with ϵ (x-axis) for XGBoost models trained over original and anonymized datasets (i.e., differentially private and k-anonymous). For each value of ϵ, the corresponding k-anonymity guarantee is $\mathbf{k} = [62, 62, 65, 70, 74]$, respectively.

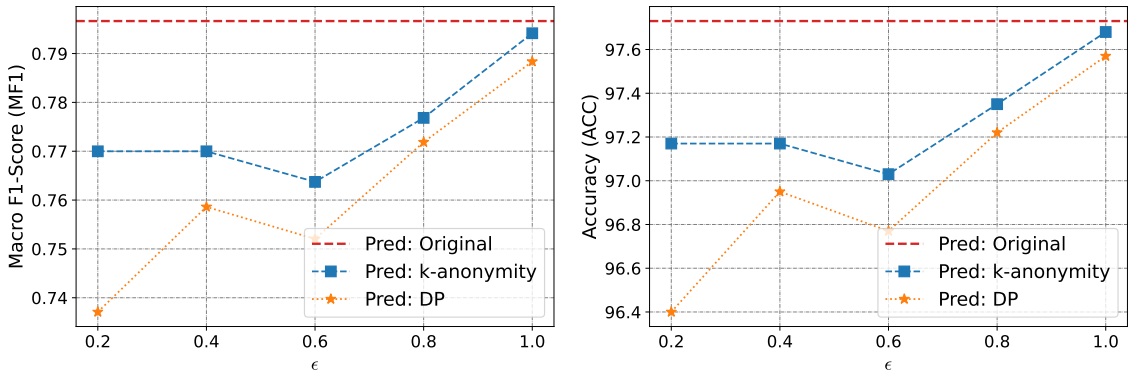

Figure 11.1: MF1 and ACC metrics (y-axis) for XGBoost models trained over original, differentially private, and k-anonymous datasets. For each value of ϵ (x-axis), the corresponding k-anonymity guarantee is $\mathbf{k} = [62, 62, 65, 70, 74]$, respectively.

One can notice from Fig. 11.1 that XGBoost models trained with anonymized data can also guarantee a good privacy-utility trade-off for the binary classification of victims' mortality. Overall, the results with the k-anonymous datasets are still close to the results with original data while the results with DP decreased more. This could be due to DP applying both sub-sampling of the learning set as well as the generalization and/or suppression of QIDs to also satisfy k-anonymity.

On the other hand, with the commonly used $\epsilon = 1$ privacy guarantee [59, 120] that also satisfies $k = 74$-anonymity, the MF1 and ACC scores of both XGBoost models trained over DP and k-anonymous data had no considerable loss of utility. Indeed, selecting $\epsilon = 1$ has also been suggested in [120] as a good parameterization value. Thus, considering $\epsilon = 1$ and $k = 74$, Table 11.1 exhibits the final generalization approach for each QID we considered of each entity (Victim – Vic. and Operator – Ope.) and privacy model (k-anonymity and DP). The symbol $*$ in Table 11.1 indicates full suppression for an attribute or masking of a digit (for Vic. City). On the one hand, the transformed/suppressed features limit the data analysis one can carry on, e.g., to find correlation between features. On the other hand, as one can notice from Fig. 11.1, although some features (the QIDs) suffered transformation and/or suppression, XGBoost models were still able to classify victims' mortality as good as with the original dataset, while providing privacy guarantees for both victims and operators.

Table 11.1: Final generalization hierarchy for each QID of each entity, namely, victim and call center operator.

Attribute	Final Generalization Hierarchy	
	k-anonymity	Differential Privacy
Vic. Age	[0, 40[, [40, 80[, [80, 101[	[0, 20[, [20, 40[, ..., [80, 101[
Vic. Gender	Feminine, Masculine, Not registered	Feminine, Masculine, Not registered, *
Vic. City	21***, 22***, 23***	212**, 222**, 233**...
Ope. Age	*	[22, 32[, [32, 42[, ..., [52, 62[
Ope. Gender	Feminine, Masculine	Feminine, Masculine, *
Ope. Seniority (in days)	[0, 2880[, [2880, 5760[, ..., [8640, 10204[	*
Ope. Grade	*	*

These results suggest that some patterns were still kept even with the transformed features. So, Fig. 11.2 illustrates 10 features with the highest impact considering the most performing XGBoost model trained over original data, $\epsilon = 1$-DP, and $k = 74$-anonymity, considering the type "Gain" feature importance algorithm. This algorithm is based on the relative contribution of each feature to improve the accuracy in the division of a branch. In Fig. 11.2, the following prefixes are used: "PROBA" for probability, "INT" for intervention, and "VIC" for victim, which corresponds to the features of the Vic_Mort-DB from Section 3.3.5.

From Fig. 11.2, we can notice that the calculated variables from probabilities (PROBA_MORT_MOT and PROBA_MORT_AGE) and the type of intervention (type, subtype, and motive) have a great impact on the creation of the models. Besides, the victims' age and gender showed more importance than the alert diffusion time (INT_D_DIFF_ALERT) and duration of the call. Lastly, in our experiments, operators' personal data did not show much importance for any XGBoost model, in this way, for upcoming works, we consider not using such predictors as there would be a need for preserving their privacy.

11.2.3/ Discussion and Related Work

As reviewed in recent survey works [164, 235, 234], several decision-support systems based on ML techniques have been proposed for application in emergency medicine. Indeed, in the context of this chapter, for EMS, there are many interests in using ML methods for tasks such as: identifying possible medical conditions before arrival on emergency departments [190], to predict ambulances' demand to allow their reallocation [75], to predict operation demand [152, 148, 193, 172, 75, 148, 176, 146, 177, 226, 154], to predict ambulance response time [215, 158], to predict the ambulances' turnaround time in hospitals [216], to predict clinical outcomes [160], to early identify clinical conditions on emergency calls [142], to recognize and predict service disruptions [178], and so on. However, to our knowledge, we are the first group investigating **privacy-preserving ML**

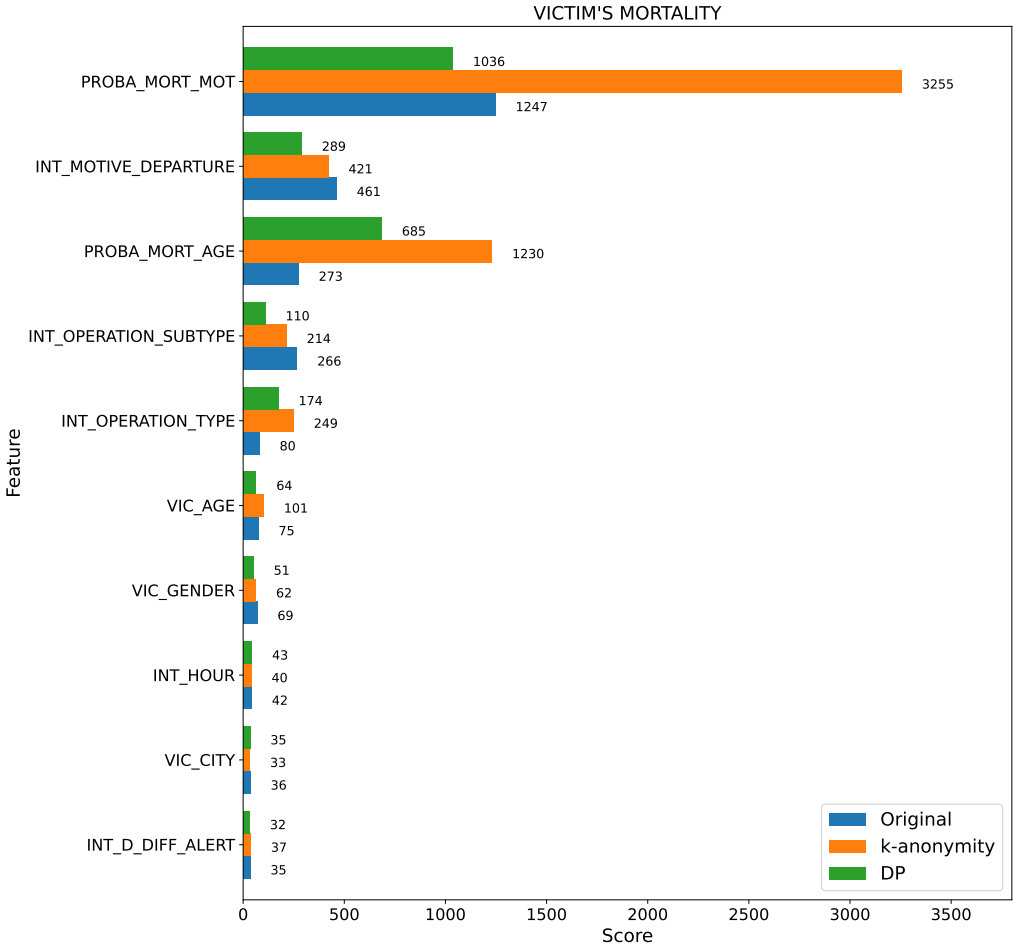

Figure 11.2: Feature importance from the XGBoost models trained over original, $\epsilon = 1$-DP, and $k = 74$-anonymity, considering the type "Gain" as score and the first 10 variables.

solutions to EMS.

Indeed, although the collection of medical data allows investigations to propose improved ML-based decision-support tools, on the other hand, there is a problem with the disclosure of personal and sensitive information. In the privacy-preserving data mining literature, there are few alternatives, e.g., objective perturbation [34], gradient perturbation [73, 241, 131], and input data perturbation [32, 31], that can help to mitigate these problems. This chapter also adopted the *input perturbation* setting (cf. Section 3.2.1) because it allows using any ML and post-processing techniques in contrast with gradient [73] or objective perturbation [34]. Furthermore, input perturbation is consistent with real-world applications in which EMS would only utilize and/or share anonymized data with third parties to train and improve ML-based decision support systems. This way, because each sample in the dataset is anonymized, data are protected from data leakage and are more difficult to reconstruct when the ML model receives attacks [105, 104, 145], for example.

11.3/ CONCLUSION

In this chapter, we aimed to predict the victims' mortality using data collected from the emergency call until an SDIS 25 center is notified about the intervention. More precisely, with all data available within this time interval (e.g., call processing times, operators' and victims' personal data, location, etc), **we sought to predict if victims will die**. This way, the SDIS 25 (or EMS in general) can quickly dispatch ambulances depending on the level of urgency. However, we also take into consideration both *victims'* and *call center operators' privacy* when training the ML models. Therefore, this interdisciplinary work aimed to evaluate the effectiveness of predicting victims' mortality with XGBoost models trained over differentially private and k-anonymous data with different levels of ϵ and k.

As shown in the results, even with anonymized datasets ($k = 74$-anonymous and $\epsilon = 1$-differentially-private), mortality could be predicted with accuracy as high as 97% with MF1 scores of about 79%. These results showed (again) the potential of privacy-preserving ML solutions for EMS, which can be used as a decision-support tool to early identify mortality while preserving the users' privacy and, thus, help EMS to save more lives. As a result of these findings, EMS may consider utilizing and/or sharing anonymised datasets to prevent data leakages, membership inference attacks, and data reconstruction attacks [228, 105, 104, 145].

Lastly, some limitations of this chapter are described in the following. First, on anonymizing the datasets, there is a clear difference in the type of privacy we provided for each *entity*. On the one hand, because *victims* were unique in our dataset, DP and k-anonymity provided *user-level* [59] privacy. On the other hand, there is a unique set of *operators* that treated many emergency calls and, thus, DP and k-anonymity provided *event-level* [59] privacy. Also, we considered an ideal case where the information of all victims in the testing set was acquired during the call. However, this may not always occur in real life, e.g., when someone activates EMS for unidentified victims.

V

CONCLUSION & PERSPECTIVES

12

CONCLUSION & PERSPECTIVES

12.1/ GENERAL CONCLUSION

In this manuscript, we approached several aspects of privacy-preserving data collection and publishing, as well as privacy-preserving machine learning. This manuscript is separated into four parts. In the **first part**, we introduced the context, the motivating projects, and the objectives. The **second part** started presenting the data privacy and ML techniques our works depend on. We finished the second part by presenting the datasets we experiment on.

The **third part** contains our contributions to privacy-preserving statistical learning, mainly with the LDP model. In the first chapter of the third part, i.e., Chapter 4, we proposed an approach to infer and recreate synthetic data that provides a precise mobility scenario based on one-week statistical data of *unions of consecutive days* made available by [53]. The generated and open dataset (https://github.com/hharcolezi/OpenMSFIMU) named MS-FIMU can be used to evaluate new privacy-preserving techniques as well as ML tasks. For instance, in Chapters 6 and 7, the MS-FIMU dataset has been used to evaluate the effectiveness of our proposed LDP protocols for multidimensional and longitudinal frequency estimates. The MS-FIMU dataset has also been used in Chapter 5, in which *we proposed an LDP-based CDRs processing system to publish multidimensional mobility reports throughout time*. We also prove in Chapter 5 that for collecting multidimensional data with GRR [80] the utility loss sending a single attribute with ϵ-LDP (i.e., *Smp* solution) is lower than splitting the privacy budget over the number of attributes. This proof extends to two other protocols named SUE [61] and OUE [108], as shown in Chapter 6.

We then abstracted the problem of Chapter 5 and thus, in Chapter 6, we focused on *improving the utility of LDP protocols for longitudinal and multidimensional frequency estimates*. Indeed, the combination of both "multi" settings (i.e., numerous attributes and longitudinal data collection) presents several problems, for which this manuscript provides the first solution called ALLOMFREE under ϵ-LDP. Under the same privacy guaran-

tee, our studies revealed that ALLOMFREE consistently and significantly outperforms the state-of-the-art protocols, namely, L-SUE (a.k.a. Basic-RAPPOR [61]) and L-OUE (i.e., OUE [108] with *memoization*), with an average accuracy increase ranging from 10% up to 55%.

We start the last chapter of the third part, i.e., Chapter 7, by arguing that the state-of-the-art *Smp* solution for multidimensional frequency estimates might be "unfair" with some users since the reported value uses the whole privacy budget ϵ, which is what is currently accepted. *We thus propose RS+FD, which may be utilized with any current LDP protocol designed for single-frequency estimation*. More precisely, with RS+FD, the client-side has two steps: local randomization and fake data generation (cf. Fig. 7.1 and Alg. 4). With our experiments, we concluded that *under the same privacy guarantee, our proposed protocols with RS+FD achieve similar or smaller estimation error than using the state-of-the-art Smp solution while enhancing all users' privacy.*

The **fourth and last part** comprises our contributions to differentially private machine learning predictions. Indeed, *the main goal of this fourth part was to evaluate the privacy-utility trade-off of training ML and DL models over differentially private data* (a.k.a. input perturbation [31, 32]). So, in Chapter 8, we developed and assessed the performance of DL models on two privacy-preserving ML settings, namely, input and gradient perturbation. For the former, we applied the Gaussian mechanism [59] to each sample before training any DL model, and for the latter, we trained DL models with the DP-SGD [73, 131, 241] algorithm. Both settings were compared on a multivariate time series forecasting task of aggregate human mobility data. *We concluded that it is still possible to have accurate multivariate forecasts in both privacy-preserving ML settings*. *In terms of accuracy* (measured with the RMSE metric), *the gradient perturbation setting surpasses input perturbation*. However, *input perturbation provides higher privacy protection than gradient perturbation* as it might also protect the aggregated mobility data against known threats (e.g., data breaches [228], membership inference attacks [103, 198], and trajectory recovery attack [135, 109]).

Next, we started to focus on our second motivating project with a collaboration with Selene Cerna, which concerns the SDIS 25 (i.e., an EMS in France). *Our assumption is that EMS intends to deploy decision-support systems to optimize their services but only shares sanitized data with the development team (i.e., third parties)*. So, in Chapter 9, *we proposed an LDP-based methodology to allow EMS to properly sanitize interventions' data*. With several experiments in frequency estimation and input perturbation-based ML forecasting, we concluded that interventions data can be properly sanitized to avoid leakage of information while remaining useful for both statistical learning (cf. Fig. 9.4) and forecasting (cf. Fig. 9.6) purposes.

Moreover, in our two last contribution chapters, i.e., Chapter 10 and 11, *we evaluated*

the privacy-utility trade-off of solutions based on ML and DP with a focus on *optimizing EMS response time to emergencies*. More precisely, in Chapter 10, we proposed to evaluate the effectiveness of several values of ϵ (i.e., the privacy budget) to sanitize emergency location data with geo-indistinguishability [46] and train ML-based models to predict ambulance response times. We concluded that the sanitization of location data with geo-indistinguishability and the perturbation of its associated features (e.g., city, distance) had a certain impact on data utility (see Table 10.2) but no considerable effect on predicting ARTs (see Fig. 10.3). Finally, in Chapter 11, we concentrate our attention on each *victim* by using several sensitive attributes of both victims and call center operators (cf. Section 3.3.5). Our objective was to use data collected within the time of the emergency call until an SDIS 25 center is notified about the intervention to predict the victims' mortality. Once more, we focused on anonymizing/sanitizing the dataset and, thus, we assessed the effectiveness of both k-anonymity [18, 20] and DP [27, 26, 59] models. That chapter concludes that even with anonymized datasets, victims' mortality could be predicted with high accuracy and macro f1-scores, which could be used as a decision-support tool by EMS to early identify high urgent situations.

12.2/ PERSPECTIVES

The research fields in privacy and privacy-preserving ML are broad and promising. For instance, it is possible and interesting to investigate the following topics in the **short term**:

- To integrate the proposed LDP protocols from Chapters 6 and 7 into the LDP-based CDRs processing system of Chapter 5 in order to improve the privacy of MNOs clients.

- To investigate how to combine the optimal longitudinal LDP protocols from Chapter 6 (i.e., L-GRR and L-OSUE) and our proposed RS+FD solution from Chapter 7 is also a planned and indicated direction.

- For both Chapters 8 and 9, for future work, we suggest and intend to investigate a more complex DL architecture to improve the results of DL/ML models proposed in this manuscript for their respective multivariate time series forecasting task.

- Regarding the work on Chapter 10, the intended future works are to extend the analysis and predictions to different operation times of EMS such as the pre-travel delay (i.e., gathering personnel and ambulances) and travel times (e.g., from the center to the emergency scene, from the emergency scene to hospitals), **while respecting users' privacy**.

In addition, for the **long term**, we list below some perspectives of the works in this manuscript:

- To extend the proposed LDP-based CDRs processing system of Chapter 5 to the shuffle DP model [141, 149, 184, 202, 224], which could provide strong privacy guarantees as well as accurate multidimensional frequency estimates (i.e., mobility reports throughout time).
- To investigate our proposed RS+FD solution from Chapter 7 on generating synthetic data from ϵ-LDP multidimensional frequency estimates for classification/regression tasks (e.g., as in [93]) in two perspectives: **performance** and **privacy-protection** (e.g., against membership inference attacks).
- To study if given a reported tuple $\mathbf{y}$ one can state which attribute value is "fake" or not by seeing the estimated frequencies reported with our RS+FD solution from Chapter 7. Indeed, we suggest investigating this phenomenon in both single-time collection and longitudinal studies (i.e., throughout time).
- Concerning multivariate time-series forecasting (i.e., Chapters 8 and 9), investigating the data leakage through membership inference attacks of both differentially private input and gradient perturbation settings is also a prospective and intended direction.
- Some future work for Chapter 11 would be to investigate a uniform notion of privacy for both entities (i.e., a set of operators linked to many unique victims). In addition, we intend to evaluate privacy-preserving ML models with randomly excluded data from victims (i.e., sex and age) since these data might not be acquired during the calls, in order to assess the models' robustness. Besides, another prospective direction would be working with the *text observations* registered by operators during calls, which could be treated with natural language processing techniques, while preserving the privacy of the individuals concerned (e.g., [230, 242]).

13

PUBLICATIONS

During the period of this thesis, the author has published the following papers and resources. The superscript * highlights equal contribution for co-first authors **in bold**.

JOURNAL PAPERS

- *__Arcolezi, H. H.__, *__Cerna, S.__, Couchot, J.-F, Guyeux, C., & Makhoul, A. (2021) Privacy-Preserving Prediction of Victim's Mortality and Their Need for Transportation to Health Facilities. **IEEE Transactions on Industrial Informatics**, Early Access [221].
- **Arcolezi, H. H.**, Cerna, S., Guyeux, C., & Couchot, J.-F. (2021). Preserving Geo-Indistinguishability of the Emergency Scene to Predict Ambulance Response Time. **Mathematical and Computational Applications**, 26(3), 56 [212].
- **Arcolezi, H. H.**, Couchot, J.-F., Cerna, S., Guyeux, C., Royer, G., Al Bouna, B., & Xiao, X. (2020). Forecasting the Number of Firefighters Interventions per Region with Local-Differential-Privacy-Based Data. **Computers & Security**, 96, 101888 [172].

CONFERENCE PAPERS

- **Arcolezi, H. H.**, Couchot, J.-F., Al Bouna, B., & Xiao, X. (2021). Random Sampling Plus Fake Data: Multidimensional Frequency Estimates With Local Differential Privacy. In Proceedings of the 30th ACM *International Conference on Information and Knowledge Management* (CIKM '21), November, Virtual Event, QLD, Australia [213].
- **Arcolezi, H. H.**, Couchot, J.-F., Al Bouna, B., & Xiao, X. (2020). Longitudinal Collection and Analysis of Mobile Phone Data with Local Differential Privacy. In Proceed-

ings of the 15th IFIP *International Summer School on Privacy and Identity Management*, September, 40-57. Springer, Cham [215].

- **Arcolezi, H. H.**, Couchot, J.-F., Baala, O., Contet, J.-M., Al Bouna, B., & Xiao, X. (2020). Mobility modeling through mobile data: generating an optimized and open dataset respecting privacy. In Proceedings of the 16th *International Wireless Communications and Mobile Computing* (IWCMC), June, 1689–1694 [171].

SUBMITTED PAPERS

- **Arcolezi, H. H.**, Couchot, J.-F., Al Bouna, B., & Xiao, X. Improving the Utility of Locally Differentially Private Protocols for Longitudinal and Multidimensional Frequency Estimates. **Digital Communications and Networks**. Submitted in August 2021 [214].

- **Arcolezi, H. H.**, Couchot, J.-F., Renaud, D., Al Bouna, B., & Xiao, X. Differentially Private Multivariate Time Series Forecasting of Aggregated Human Mobility With Deep Learning: Input or Gradient Perturbation? **Neural Computing and Applications**. Submitted in September 2021.

CO-AUTHORED PAPERS

Furthermore, the author also participated as a co-author in the following published papers.

- Cerna, S., **Arcolezi, H. H.**, Guyeux, C., Royer-Fey, G., & Chevallier, C. (2021). Machine learning-based forecasting of firemen ambulances' turnaround time in hospitals, considering the COVID-19 impact. **Applied Soft Computing**, 109, 107561 [216].

- Cisneros, L. L., **Arcolezi, H. H.**, Cerna, S., Brandão, J.L., Santos, G.C., Navarro, T.P., & Carvalho, A.A. (2021). Machine Learning Algorithms to Predict In-Hospital Mortality in Patients with Diabetic Foot Ulceration. In Proceedings of the *XXIII Congresso da Sociedade Brasileira de Diabetes*.

- Cerna, S., Guyeux, C., **Arcolezi, H. H.**, Couturier, R., & Royer, G. (2020). A comparison of LSTM and XGBoost for predicting firemen interventions. In Proceedings of the 8th *World Conference on Information Systems and Technologies* (WorldCIST), April, 424–434 [176].

- Cerna, S., Guyeux, C., **Arcolezi, H. H.**, & Royer, G. (2020). Boosting Methods for Predicting Firemen Interventions. In Proceedings of the 11th *International Conference on Information and Communication Systems* (ICICS), 001–006 [177].

RESOURCES & CODES

The generated MS-FIMU dataset of Chapter 4 is fully available on the following GitHub page:

• https://github.com/hharcolezi/OpenMSFIMU.

Lastly, the author also maintains a list of DP and LDP experiments of the work carried out in Chapters 5, 6, 7, 8, and 10 on the following GitHub page:

• https://github.com/hharcolezi/ldp-protocols-mobility-cdrs.

BIBLIOGRAPHY

[1] **Bias of an estimator**. Available online: https://en.wikipedia.org/wiki/Bias_of_an_estimator (accessed on 15 October 2021).

[2] **Confinements liés à la pandémie de COVID-19 en france**. Available online: https://fr.wikipedia.org/wiki/Confinements_li%C3%A9s_%C3%A0_la_pand%C3%A9mie_de_Covid-19_en_France (accessed on 11 July 2021).

[3] **Great-circle distance**. Available online: https://en.wikipedia.org/wiki/Great-circle_distance (accessed on 05 October 2021).

[4] **Interquartile range**. Available online: https://en.wikipedia.org/wiki/Interquartile_range (accessed on 16 October 2021).

[5] **Lambert w function**. Available online: https://en.wikipedia.org/wiki/Lambert_W_function (accessed on 29 September 2021).

[6] **Location guard**. https://github.com/chatziko/location-guard.

[7] **Pearson correlation coefficient**. Available online: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient (accessed on 15 October 2021).

[8] **Seattle fire department: Real-time 911 dispatch**. Available online: http://www2.seattle.gov/fire/realtime911/ (accessed on 18 February 2021).

[9] **Service Départemental d'Incendie et de Secours du Doubs (SDIS 25)**. Available online: https://www.sdis25.fr/ (accessed on 10 October 2021).

[10] **Universal declaration of human rights**, 1948. Available online: https://www.un.org/en/about-us/universal-declaration-of-human-rights (accessed on 01 October 2021).

[11] WARNER, S. L. **Randomized response: A survey technique for eliminating evasive answer bias**. *Journal of the American Statistical Association 60*, 309 (Mar. 1965), 63–69.

[12] DALENIUS, T. **Towards a methodology for statistical disclosure control**. *statistik Tidskrift 15*, 429-444 (1977), 2–1.

[13] **Commission nationale de l'informatique et des libertés (CNIL)**, 1978. Available online: https://www.cnil.fr/en/home (accessed on 04 July 2021).

[14] LEWIS, C. **Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting**. Butterworth scientific. Butterworth Scientific, 1982.

[15] TIBSHIRANI, R. **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 1 (1996), 267–288.

[16] HOCHREITER, S., AND SCHMIDHUBER, J. **Long short-term memory**. *Neural computation 9*, 8 (1997), 1735–1780.

[17] SCHUSTER, M., AND PALIWAL, K. **Bidirectional recurrent neural networks**. *IEEE Transactions on Signal Processing 45*, 11 (1997), 2673–2681.

[18] SAMARATI, P., AND SWEENEY, L. **Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression**.

[19] PONS, P. T., AND MARKOVCHICK, V. J. **Eight minutes or less: does the ambulance response time guideline impact trauma patient outcome?** *The Journal of Emergency Medicine 23*, 1 (July 2002), 43–48.

[20] SWEENEY, L. **k-anonymity: A model for protecting privacy**. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10*, 05 (Oct. 2002), 557–570.

[21] AUSTIN, P. C. **Quantile regression: A statistical tool for out-of-hospital research**. *Academic Emergency Medicine 10*, 7 (July 2003), 789–797.

[22] BRODER, A., AND MITZENMACHER, M. **Network applications of bloom filters: A survey**. *Internet Mathematics 1*, 4 (Jan. 2004), 485–509.

[23] PELEG, K., AND PLISKIN, J. S. **A geographic information system simulation model of EMS: reducing ambulance response time**. *The American Journal of Emergency Medicine 22*, 3 (May 2004), 164–170.

[24] BARBARO, M., AND JR., T. Z. **A face is exposed for aol searcher no. 4417749**, 2006. Available online: https://www.nytimes.com/2006/08/09/technology/09aol.html (accessed on 14 October 2021).

[25] CHAUDHURI, K., AND MISHRA, N. **When random sampling preserves privacy**. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006, pp. 198–213.

[26] DWORK, C. **Differential privacy**. In *Automata, Languages and Programming*. Springer Berlin Heidelberg, 2006, pp. 1–12.

[27] DWORK, C., MCSHERRY, F., NISSIM, K., AND SMITH, A. **Calibrating noise to sensitivity in private data analysis**. In *Theory of Cryptography*. Springer Berlin Heidelberg, 2006, pp. 265–284.

[28] MACHANAVAJJHALA, A., GEHRKE, J., KIFER, D., AND VENKITASUBRAMANIAM, M. **L-diversity: privacy beyond k-anonymity**. In *22nd International Conference on Data Engineering (ICDE'06)* (2006), IEEE.

[29] LI, N., LI, T., AND VENKATASUBRAMANIAN, S. **t-closeness: Privacy beyond k-anonymity and l-diversity**. In *2007 IEEE 23rd International Conference on Data Engineering* (Apr. 2007), IEEE.

[30] SILVERMAN, R. A., GALEA, S., BLANEY, S., FREESE, J., PREZANT, D. J., PARK, R., PAHK, R., CARON, D., YOON, S., EPSTEIN, J., AND RICHMOND, N. J. **The "vertical response time": Barriers to ambulance response in an urban area**. *Academic Emergency Medicine 14*, 9 (Sept. 2007), 772–778.

[31] AGGARWAL, C. C., AND YU, P. S., Eds. **Privacy-Preserving Data Mining**. Springer US, 2008.

[32] KASIVISWANATHAN, S. P., LEE, H. K., NISSIM, K., RASKHODNIKOVA, S., AND SMITH, A. **What can we learn privately?** In *2008 49th Annual IEEE Symposium on Foundations of Computer Science* (Oct. 2008), IEEE.

[33] ALADDINI, K. **EMS response time models: A case study and analysis for the region of waterloo**. Master's thesis, University of Waterloo, 2010.

[34] CHAUDHURI, K., MONTELEONI, C., AND SARWATE, A. D. **Differentially private empirical risk minimization.** *Journal of Machine Learning Research 12*, 3 (2011).

[35] LUXEN, D., AND VETTER, C. **Real-time routing with openstreetmap data**. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (New York, NY, USA, 2011), GIS '11, ACM, pp. 513–516.

[36] PEDREGOSA, F., AND OTHERS. **Scikit-learn: Machine learning in Python**. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[37] SHOKRI, R., THEODORAKOPOULOS, G., BOUDEC, J.-Y. L., AND HUBAUX, J.-P. **Quantifying location privacy**. In *2011 IEEE Symposium on Security and Privacy* (May 2011), IEEE.

[38] ZANG, H., AND BOLOT, J. **Anonymization of location data does not work**. In *Proceedings of the 17th annual international conference on Mobile computing and networking - MobiCom* (2011), ACM Press.

[39] ABOUELJINANE, L., JEMAI, Z., AND SAHIN, E. **Reducing ambulance response time using simulation: The case of val-de-marne department emergency medical service**. In *Proceedings Title: Proceedings of the 2012 Winter Simulation Conference (WSC)* (Dec. 2012), IEEE.

[40] BERGSTRA, J., AND BENGIO, Y. **Random search for hyper-parameter optimization.** *Journal of machine learning research 13*, 2 (2012).

[41] DO, Y. K., FOO, K., NG, Y. Y., AND ONG, M. E. H. **A quantile regression analysis of ambulance response time**. *Prehospital Emergency Care 17*, 2 (Dec. 2012), 170–176.

[42] FARAGLIA, D. **Faker**, 2012. Available online: https://github.com/joke2k/faker (accessed on 25 September 2019).

[43] LI, N., QARDAJI, W., AND SU, D. **On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy**. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security - ASIACCS '12* (2012), ACM Press.

[44] LU, X., BENGTSSON, L., AND HOLME, P. **Predictability of population displacement after the 2010 haiti earthquake**. *Proceedings of the National Academy of Sciences 109*, 29 (June 2012), 11576–11581.

[45] **Statistiques mensuelles fournies par le service départemental d'incendies et de secours (sdis 71)**, 2013. Available online: https://www.data.gouv.fr/fr/datasets/interventions-des-pompiers-od71/ (accessed on 13 December 2019).

[46] ANDRÉS, M. E., BORDENABE, N. E., CHATZIKOKOLAKIS, K., AND PALAMIDESSI, C. **Geo-indistinguishability**. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13* (2013), ACM Press.

[47] BERGSTRA, J., YAMINS, D., AND COX, D. D. **Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures**. In *Proceedings of the 30th International Conference on International Conference on Machine Learning* (2013), ICML'13, JMLR, p. I–115–I–123.

[48] CHATZIKOKOLAKIS, K., ANDRÉS, M. E., BORDENABE, N. E., AND PALAMIDESSI, C. **Broadening the scope of differential privacy using metrics**. In *Privacy Enhancing Technologies*. Springer Berlin Heidelberg, 2013, pp. 82–102.

[49] DE MONTJOYE, Y.-A., HIDALGO, C. A., VERLEYSEN, M., AND BLONDEL, V. D. **Unique in the crowd: The privacy bounds of human mobility**. *Scientific Reports 3*, 1 (Mar. 2013).

[50] DUCHI, J. C., JORDAN, M. I., AND WAINWRIGHT, M. J. **Local privacy and statistical minimax rates**. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science* (Oct. 2013), IEEE.

[51] JAMES, G., WITTEN, D., HASTIE, T., AND TIBSHIRANI, R. **An introduction to statistical learning**, vol. 112. Springer, 2013.

[52] MIR, D. J., ISAACMAN, S., CACERES, R., MARTONOSI, M., AND WRIGHT, R. N. **DP-WHERE: Differentially private modeling of human mobility**. In *2013 IEEE International Conference on Big Data* (Oct. 2013), IEEE.

[53] ORANGE-BUSINESS-SERVICES. **Flux vision: real time statistics on mobility patterns**, 2013. Available online: https://www.orange-business.com/en/products/flux-vision (accessed on 05 October 2021).

[54] SARWATE, A. D., AND CHAUDHURI, K. **Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data**. *IEEE Signal Processing Magazine 30*, 5 (Sept. 2013), 86–94.

[55] ACS, G., AND 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 '14* (2014), ACM Press.

[56] BUCKEE, C. O. **Protect privacy of mobile data**. *Nature 514*, 7520 (Oct. 2014), 35–35.

[57] CHUNG, J., GULCEHRE, C., CHO, K., AND BENGIO, Y. **Empirical evaluation of gated recurrent neural networks on sequence modeling**. In *NIPS 2014 Workshop on Deep Learning* (2014).

[58] DREDGE, S. **Tinder dating app was sharing more of users' location data than they realised**, 2014. Available online: https://www.theguardian.com/technology/2014/feb/20/tinder-app-dating-data-location-sharing (accessed on 14 October 2021).

[59] DWORK, C., ROTH, A., AND OTHERS. **The algorithmic foundations of differential privacy**. *Foundations and Trends® in Theoretical Computer Science 9*, 3–4 (2014), 211–407.

[60] DWORK, C., TALWAR, K., THAKURTA, A., AND ZHANG, L. **Analyze gauss: Optimal bounds for privacy-preserving principal component analysis**. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing* (New York, NY, USA, 2014), STOC '14, Association for Computing Machinery, p. 11–20.

[61] ERLINGSSON, U., PIHUR, V., AND KOROLOVA, A. **RAPPOR: Randomized aggregatable privacy-preserving ordinal response**. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2014), ACM, pp. 1054–1067.

[62] HEERSCHAP, N., ORTEGA, S., PRIEM, A., AND OFFERMANS, M. **Innovation of tourism statistics through the use of new big data sources**. In *12th global forum on tourism statistics, Prague, CZ* (2014), vol. 716.

[63] HSU, J., GABOARDI, M., HAEBERLEN, A., KHANNA, S., NARAYAN, A., PIERCE, B. C., AND ROTH, A. **Differential privacy: An economic method for choosing epsilon**. In *Proceedings of the 2014 IEEE 27th Computer Security Foundations Symposium* (Washington, DC, USA, 2014), CSF '14, IEEE Computer Society, pp. 398–410.

[64] TROTTER, J. **Public nyc taxicab database lets you see how celebrities tip**, 2014. Available online: https://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546 (accessed on 14 October 2021).

[65] WESOLOWSKI, A., BUCKEE, C. O., BENGTSSON, L., WETTER, E., LU, X., AND TATEM, A. J. **Commentary: Containing the ebola outbreak - the potential and challenge of mobile network data**. *PLoS Currents* (2014).

[66] ALAGGAN, M., GAMBS, S., MATWIN, S., AND TUHIN, M. **Sanitization of call detail records via differentially-private bloom filters**. In *Data and Applications Security and Privacy XXIX*. Springer International Publishing, 2015, pp. 223–230.

[67] BASSILY, R., AND SMITH, A. **Local, private, efficient protocols for succinct histograms**. In *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing* (June 2015), ACM.

[68] CHOLLET, F., AND OTHERS. **Keras**. https://keras.io, 2015.

[69] LECUN, Y., BENGIO, Y., AND HINTON, G. **Deep learning**. *Nature 521*, 7553 (May 2015), 436–444.

[70] PRASSER, F., AND KOHLMAYER, F. **Putting statistical disclosure control into practice: The ARX data anonymization tool**. In *Medical Data Privacy Handbook*. Springer International Publishing, 2015, pp. 111–148.

[71] SHOKRI, R., AND SHMATIKOV, V. **Privacy-preserving deep learning**. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Oct. 2015), ACM.

[72] SWEENEY, L. **Only you, your doctor, and many others may know**. *Technology Science 2015092903*, 9 (2015), 29.

[73] ABADI, M., CHU, A., GOODFELLOW, I., MCMAHAN, H. B., MIRONOV, I., TALWAR, K., AND ZHANG, L. **Deep learning with differential privacy**. CCS '16, Association for Computing Machinery, p. 308–318.

[74] CAIATI, V., BEDOGNI, L., BONONI, L., FERRERO, F., FIORE, M., AND VESCO, A. **Estimating urban mobility with open data: A case study in bologna**. In *2016 IEEE International Smart Cities Conference (ISC2)* (Sept. 2016), IEEE.

[75] CHEN, A. Y., LU, T.-Y., MA, M. H.-M., AND SUN, W.-Z. **Demand forecast using data analytics for the preallocation of ambulances**. *IEEE Journal of Biomedical and Health Informatics 20*, 4 (July 2016), 1178–1187.

[76] CHEN, T., AND GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2016), ACM.

[77] ELSALAMOUNY, E., AND GAMBS, S. **Differential Privacy Models for Location-Based Services**. *Transactions on Data Privacy 9*, 1 (2016), 15 – 48.

[78] FANTI, G., PIHUR, V., AND ERLINGSSON, Ú. **Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries**. *Proceedings on Privacy Enhancing Technologies 2016*, 3 (May 2016), 41–61.

[79] GOODFELLOW, I., BENGIO, Y., COURVILLE, A., AND BENGIO, Y. **Deep learning**, vol. 1. MIT press Cambridge, 2016.

[80] KAIROUZ, P., BONAWITZ, K., AND RAMAGE, D. **Discrete distribution estimation under local privacy**. In *International Conference on Machine Learning* (2016), PMLR, pp. 2436–2444.

[81] KAIROUZ, P., OH, S., AND VISWANATH, P. **Extremal mechanisms for local differential privacy**. *The Journal of Machine Learning Research 17*, 1 (2016), 492–542.

[82] NEHME, Z., ANDREW, E., AND SMITH, K. **Factors influencing the timeliness of emergency medical service response to time critical emergencies**. *Prehospital Emergency Care 20*, 6 (Aug. 2016), 783–791.

[83] NGUYÊN, T. T., XIAO, X., YANG, Y., HUI, S. C., SHIN, H., AND SHIN, J. **Collecting and analyzing data from smart device users with local differential privacy.** *ArXiv abs/1606.05053* (2016).

[84] PENN, C., KOOLE, T., AND NATTRASS, R. **When seconds count: A study of communication variables in the opening segment of emergency calls**. *Journal of Health Psychology 22*, 10 (Feb. 2016), 1256–1264.

[85] QIN, Z., YANG, Y., YU, T., KHALIL, I., XIAO, X., AND REN, K. **Heavy hitter estimation over set-valued data with local differential privacy**. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Oct. 2016), ACM.

[86] SHAHRIARI, B., SWERSKY, K., WANG, Z., ADAMS, R. P., AND DE FREITAS, N. **Taking the human out of the loop: A review of bayesian optimization**. 148–175.

[87] ALAGGAN, M., CUNCHE, M., AND MINIER, M. **Non-interactive (t, n)-incidence counting from differentially private indicator vectors**. In *Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics* (Mar. 2017), ACM.

[88] BASSILY, R., NISSIM, K., STEMMER, U., AND THAKURTA, A. **Practical locally private heavy hitters**. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2017), NIPS'17, Curran Associates Inc., p. 2285–2293.

[89] BITTAU, A., ERLINGSSON, Ú., MANIATIS, P., MIRONOV, I., RAGHUNATHAN, A., LIE, D., RUDOMINER, M., KODE, U., TINNES, J., AND SEEFELD, B. **Prochlo: Strong privacy for analytics in the crowd**. In *Proceedings of the 26th Symposium on Operating Systems Principles* (Oct. 2017), ACM.

[90] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., AND STONE, C. J. **Classification And Regression Trees**. Routledge, Oct. 2017.

[91] CHATZIKOKOLAKIS, K., ELSALAMOUNY, E., PALAMIDESSI, C., AND PAZII, A. **Methods for location privacy: A comparative overview**. *Foundations and Trends® in Privacy and Security 1*, 4 (2017), 199–257.

[92] CHOLLET, F. **Deep Learning with Python**, 1st ed. Manning Publications Co., USA, 2017.

[93] CYPHERS, B., AND VEERAMACHANENI, K. **AnonML: Locally private machine learning over a network of peers**. IEEE.

[94] DANCOURT, A.-C. **FIMU belfort 2017 : le festival parfait pour bouger à la pentecôte**, 2017. Available online: http://www.leparisien.fr/culture-loisirs/fimu-belfort-2017-le-festival-parfait-pour-bouger-a-la-pentecote-23-05-2017-6976476.php (accessed on 05 November 2019).

[95] DING, B., KULKARNI, J., AND YEKHANIN, S. **Collecting telemetry data privately**. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3571–3580.

[96] DUA, D., AND GRAFF, C. **UCI machine learning repository**, 2017.

[97] DUPORTAIL, J. **I asked tinder for my data. it sent me 800 pages of my deepest, darkest secrets**, 2017. Available online: https://www.theguardian.com/technology/2017/sep/26/tinder-personal-data-dating-app-messages-hacked-sold (accessed on 14 October 2021).

[98] FUKUCHI, K., TRAN, Q. K., AND SAKUMA, J. **Differentially private empirical risk minimization with input perturbation**. In *Discovery Science*. Springer International Publishing, 2017, pp. 82–90.

[99] KASHIYAMA, T., PANG, Y., AND SEKIMOTO, Y. **Open PFLOW: Creation and evaluation of an open dataset for typical people mass movement in urban areas**. *Transportation Research Part C: Emerging Technologies 85* (Dec. 2017), 249–267.

[100] KE, G., MENG, Q., FINLEY, T., WANG, T., CHEN, W., MA, W., YE, Q., AND LIU, T.-Y. **Lightgbm: A highly efficient gradient boosting decision tree**. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3146–3154.

[101] PAPPALARDO, L., AND SIMINI, F. **Data-driven generation of spatio-temporal routines in human mobility**. *Data Mining and Knowledge Discovery 32*, 3 (Dec. 2017), 787–829.

[102] PHAN, N., WU, X., HU, H., AND DOU, D. **Adaptive laplace mechanism: Differential privacy preservation in deep learning**. In *2017 IEEE International Conference on Data Mining (ICDM)* (2017), pp. 385–394.

[103] PYRGELIS, A., TRONCOSO, C., AND CRISTOFARO, E. D. **What does the crowd say about you? evaluating aggregation-based location privacy**. *Proceedings on Privacy Enhancing Technologies 2017*, 4 (Oct. 2017), 156–176.

[104] SHOKRI, R., STRONATI, M., SONG, C., AND SHMATIKOV, V. **Membership inference attacks against machine learning models**. In *2017 IEEE Symposium on Security and Privacy (SP)* (May 2017), IEEE.

[105] SONG, C., RISTENPART, T., AND SHMATIKOV, V. **Machine learning models that remember too much**. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Oct. 2017), ACM.

[106] TEAM, A. D. P. **Learning with privacy at scale**, dec 2017. Available online: https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf (accessed on 11 March 2021).

[107] WANG, S., NIE, Y., WANG, P., XU, H., YANG, W., AND HUANG, L. **Local private ordinal data distribution estimation**. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications* (May 2017), IEEE.

[108] WANG, T., BLOCKI, J., LI, N., AND JHA, S. **Locally differentially private protocols for frequency estimation**. In *26th USENIX Security Symposium (USENIX Security 17)* (Vancouver, BC, Aug. 2017), USENIX Association, pp. 729–745.

[109] XU, F., TU, Z., LI, Y., ZHANG, P., FU, X., AND JIN, D. **Trajectory recovery from ash**. In *Proceedings of the 26th International Conference on World Wide Web* (Apr. 2017), International World Wide Web Conferences Steering Committee.

[110] ZHU, T., LI, G., ZHOU, W., AND YU, P. S. **Differential Privacy and Applications**. Springer International Publishing, 2017.

[111] **Données hebdomadaires sur les interventions des sapeurs-pompiers de l'essonne**, 2018. Available online: https://www.data.gouv.fr/fr/datasets/interventions-des-pompiers/ (accessed on 13 December 2019).

[112] **General data protection regulation (GDPR)**, 2018. Available online: https://gdpr-info.eu/ (accessed on 04 July 2021).

[113] **Liste et composition 2018**, 2018. Available online: https://www.collectivites-locales.gouv.fr/liste-et-composition-2018/ (accessed on 01 December 2019).

[114] ABOWD, J. M. **The U.S. census bureau adopts differential privacy**. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (July 2018), ACM.

[115] ALAGGAN, M., CUNCHE, M., AND GAMBS, S. **Privacy-preserving Wi-Fi Analytics**. *Proceedings on Privacy Enhancing Technologies 2018*, 2 (2018), 4–26.

[116] ALVIM, M., CHATZIKOKOLAKIS, K., PALAMIDESSI, C., AND 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)* (July 2018), IEEE.

[117] ALVIM, M. S., CHATZIKOKOLAKIS, K., PALAMIDESSI, C., AND PAZII, A. **Metric-based local differential privacy for statistical applications**. In *31st Computer Security Foundations Symposium (CSF 2018)* (Oxford, United Kingdom, Jul 2018), IEEE Computer Society, pp. 262–267.

[118] BALLE, B., BARTHE, G., AND GABOARDI, M. **Privacy amplification by subsampling: tight analyses via couplings and divergences**. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (2018), pp. 6280–6290.

[119] BEAL, L., HILL, D., MARTIN, R., AND HEDENGREN, J. **Gekko optimization suite**. *Processes 6*, 8 (2018), 106.

[120] BILD, R., KUHN, K. A., AND PRASSER, F. **SafePub: A truthful data anonymization algorithm with strong privacy guarantees**. *Proceedings on Privacy Enhancing Technologies 2018*, 1 (Jan. 2018), 67–87.

[121] BÜRGER, A., WNENT, J., BOHN, A., JANTZEN, T., BRENNER, S., LEFERING, R., SEEWALD, S., GRÄSNER, J.-T., AND FISCHER, M. **The effect of ambulance response time on survival following out-of-hospital cardiac arrest**. *Deutsches Aerzteblatt Online* (Aug. 2018).

[122] DE MONTJOYE, Y.-A., AND OTHERS. **On the privacy-conscientious use of mobile phone data**. *Scientific Data 5*, 1 (Dec. 2018).

[123] DUCHI, J. C., JORDAN, M. I., AND WAINWRIGHT, M. J. **Minimax optimal procedures for locally private estimation**. *Journal of the American Statistical Association 113*, 521 (Jan. 2018), 182–201.

[124] FAN, L. **Image pixelization with differential privacy**. Springer International Publishing, 2018, pp. 148–162.

[125] FELDMAN, V., MIRONOV, I., TALWAR, K., AND THAKURTA, A. **Privacy amplification by iteration**. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)* (2018), pp. 521–532.

[126] HERN, A. **Uber fined £385,000 for data breach affecting millions of passengers**, 2018. Available online: https://www.theguardian.com/technology/2018/nov/27/uber-fined-385000-for-data-breach-affecting-millions-of-passengers-hacked (accessed on 14 October 2021).

[127] HONG, L., LEE, M., MASHHADI, A., AND FRIAS-MARTINEZ, V. **Towards understanding communication behavior changes during floods using cell phone data**. In *Lecture Notes in Computer Science*. Springer International Publishing, 2018, pp. 97–107.

[128] HYNDMAN, R. J., AND ATHANASOPOULOS, G. **Forecasting: principles and practice**. OTexts, 2018.

[129] KIM, J. W., KIM, D.-H., AND JANG, B. **Application of local differential privacy to collection of indoor positioning data**. *IEEE Access 6* (2018), 4276–4286.

[130] KONG, X., XIA, F., NING, Z., RAHIM, A., CAI, Y., GAO, Z., AND MA, J. **Mobility dataset generation for vehicular social networks based on floating car data**. *IEEE Transactions on Vehicular Technology 67*, 5 (May 2018), 3874–3886.

[131] MCMAHAN, H. B., ANDREW, G., ERLINGSSON, U., CHIEN, S., MIRONOV, I., PAPERNOT, N., AND KAIROUZ, P. **A general approach to adding differential privacy to iterative training procedures**. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop on Privacy Preserving Machine Learning* (2018).

[132] NEAR, J. **Differential privacy at scale: Uber and berkeley collaboration**. In *Enigma 2018 (Enigma 2018)* (Santa Clara, CA, Jan. 2018), USENIX Association.

[133] OUYANG, K., SHOKRI, R., ROSENBLUM, D. S., AND YANG, W. **A non-parametric generative model for human trajectories**. IJCAI'18, AAAI Press, p. 3812–3817.

[134] REN, X., YU, C.-M., YU, W., YANG, S., MEMBER, S., YANG, X., MCCANN, J. A., YU, P. S., AND FELLOW, L. **LoPub : High-Dimensional Crowdsourced Data**. 2151–2166.

[135] TU, Z., XU, F., LI, Y., ZHANG, P., AND JIN, D. **A new privacy breach: User trajectory recovery from aggregated mobility data**. *IEEE/ACM Transactions on Networking 26*, 3 (June 2018), 1446–1459.

[136] WANG, T., LI, N., AND JHA, S. **Locally differentially private frequent itemset mining**. In *2018 IEEE Symposium on Security and Privacy (SP)* (May 2018), IEEE.

[137] YIN, M., SHEEHAN, M., FEYGIN, S., PAIEMENT, J.-F., AND POZDNOUKHOV, A. **A generative model of urban activities from cellular data**. *IEEE Transactions on Intelligent Transportation Systems 19*, 6 (2018), 1682–1696.

[138] ZHANG, Z., WANG, T., LI, N., HE, S., AND CHEN, J. **CALM: Consistent adaptive local marginal for marginal release under local differential privacy**. *Proceedings of the ACM Conference on Computer and Communications Security* (2018), 212–229.

[139] ACHARYA, J., SUN, Z., AND ZHANG, H. **Hadamard response: Estimating distributions privately, efficiently, and with little communication**. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (16–18 Apr 2019), K. Chaudhuri and M. Sugiyama, Eds., vol. 89 of *Proceedings of Machine Learning Research*, PMLR, pp. 1120–1129.

[140] AL-RUBAIE, M., AND CHANG, J. M. **Privacy-preserving machine learning: Threats and solutions**. *IEEE Security Privacy 17*, 2 (2019), 49–58.

[141] BALLE, B., BELL, J., GASCÓN, A., AND NISSIM, K. **The privacy blanket of the shuffle model**. In *Advances in Cryptology – CRYPTO 2019*. Springer International Publishing, 2019, pp. 638–667.

[142] BLOMBERG, S. N., AND OTHERS. **Machine learning as a supportive tool to recognize cardiac arrest in emergency calls**. *Resuscitation 138* (May 2019), 322–329.

[143] BUN, M., NELSON, J., AND STEMMER, U. **Heavy hitters and the structure of local privacy**. *ACM Transactions on Algorithms 15*, 4 (Oct. 2019), 1–40.

[144] BYRNE, J. P., MANN, N. C., DAI, M., MASON, S. A., KARANICOLAS, P., RIZOLI, S., AND NATHENS, A. B. **Association between emergency medical service response time and motor vehicle crash mortality in the united states**. *JAMA Surgery 154*, 4 (Apr. 2019), 286.

[145] CARLINI, N., LIU, C., ERLINGSSON, Ú., KOS, J., AND SONG, D. **The secret sharer: Evaluating and testing unintended memorization in neural networks**. In *28th USENIX Security Symposium (USENIX Security 19)* (Santa Clara, CA, Aug. 2019), USENIX Association, pp. 267–284.

[146] CERNA, S., GUYEUX, C., ARCOLEZI, H. H., LOTUFO, A. D. P., COUTURIER, R., AND ROYER, G. **Long short-term memory for predicting firemen interventions**. In *6th International Conference on Control, Decision and Information Technologies (CoDIT 2019)* (Paris, France, apr 2019).

[147] CORMODE, G., KULKARNI, T., AND SRIVASTAVA, D. **Answering range queries under local differential privacy**. *Proceedings of the VLDB Endowment 12*, 10 (June 2019), 1126–1138.

[148] COUCHOT, J.-F., GUYEUX, C., AND ROYER, G. **Anonymously forecasting the number and nature of firefighting operations**. In *Proceedings of the 23rd International Database Applications & Engineering Symposium on - IDEAS '19* (2019), ACM Press.

[149] ERLINGSSON, Ú., FELDMAN, V., MIRONOV, I., RAGHUNATHAN, A., TALWAR, K., AND THAKURTA, A. **Amplification by shuffling: From local to central differential privacy via anonymity**. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Jan. 2019, pp. 2468–2479.

[150] FERNANDES, N., LEFKI, K., AND PALAMIDESSI, C. **Utility-Preserving Privacy Mechanisms for Counting Queries**. Springer International Publishing, Cham, 2019, pp. 487–495.

[151] GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems**. O'Reilly Media, 2019.

[152] GREKOUSIS, G., AND LIU, Y. **Where will the next emergency event occur? predicting ambulance demand in emergency medical services using artificial intelligence**. *Computers, Environment and Urban Systems 76* (July 2019), 110–122.

[153] GU, X., LI, M., CAO, Y., AND XIONG, L. **Supporting both range queries and frequency estimation with local differential privacy**. In *2019 IEEE Conference on Communications and Network Security (CNS)* (June 2019), IEEE.

[154] GUYEUX, C., NICOD, J.-M., VARNIER, C., MASRY, Z. A., ZERHOUNY, N., OMRI, N., AND ROYER, G. **Firemen prediction by using neural networks: A real case study**. In *Advances in Intelligent Systems and Computing*. Springer International Publishing, Aug. 2019, pp. 541–552.

[155] JAYARAMAN, B., AND EVANS, D. **Evaluating differentially private machine learning in practice**. In *28th USENIX Security Symposium (USENIX Security 19)* (Santa Clara, CA, Aug. 2019), USENIX Association, pp. 1895–1912.

[156] KISHORE, N., MITCHELL, R., LASH, T. L., REED, C., DANON, L., SIGMUNDSDÓTTIR, G., AND 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 (Nov. 2019), 37–45.

[157] LEE, D. W., MOON, H. J., AND HEO, N. H. **Association between ambulance response time and neurologic outcome in patients with cardiac arrest**. *The American Journal of Emergency Medicine 37*, 11 (Nov. 2019), 1999–2003.

[158] LIAN, X., MELANCON, S., PRESTA, J.-R., REEVESMAN, A., SPIERING, B., AND WOODBRIDGE, D. **Scalable real-time prediction and analysis of san francisco**

fire department response times. In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (Aug. 2019), IEEE.

[159] MURAKAMI, T., AND KAWAMOTO, Y. **Utility-Optimized local differential privacy mechanisms for distribution estimation**. In *28th USENIX Security Symposium (USENIX Security 19)* (Santa Clara, CA, Aug. 2019), USENIX Association, pp. 1877–1894.

[160] MYOUNG KWON, J., AND OTHERS. **Deep-learning-based out-of-hospital cardiac arrest prognostic system to predict clinical outcomes**. *Resuscitation 139* (June 2019), 84–91.

[161] PENG, F., TANG, S., ZHAO, B., AND LIU, Y. **A privacy-preserving data aggregation of mobile crowdsensing based on local differential privacy**. In *Proceedings of the ACM Turing Celebration Conference - China* (May 2019), ACM.

[162] PIRKLBAUER, K., AND FINDLING, R. D. **Predicting the category of fire department operations**. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services* (Dec. 2019), ACM.

[163] PRIMAULT, V., BOUTET, A., MOKHTAR, S. B., AND BRUNIE, L. **The long road to computational location privacy: A survey**. *IEEE Communications Surveys & Tutorials 21*, 3 (2019), 2772–2793.

[164] SHAFAF, N., AND MALEK, H. **Applications of machine learning approaches in emergency medicine; a review article**. *Archives of Academic Emergency Medicine 7*, 1 (July 2019).

[165] SOYKAN, E. U., BILGIN, Z., ERSOY, M. A., AND TOMUR, E. **Differentially private deep learning for load forecasting on smart grid**. In *2019 IEEE Globecom Workshops (GC Wkshps)* (Dec. 2019), IEEE.

[166] WANG, N., XIAO, X., YANG, Y., ZHAO, J., HUI, S. C., SHIN, H., SHIN, J., AND YU, G. **Collecting and analyzing multidimensional data with local differential privacy**. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)* (Apr. 2019), IEEE.

[167] WONG, J. C. **Facebook to be fined $5bn for cambridge analytica privacy violations – reports**, 2019. Available online: https://www.theguardian.com/technology/2019/jul/12/facebook-fine-ftc-privacy-violations (accessed on 14 October 2021).

[168] ZHAO, D., CHEN, H., ZHAO, S., ZHANG, X., LI, C., AND LIU, R. **Local differential privacy with k-anonymous for frequency estimation**. In *2019 IEEE International Conference on Big Data (Big Data)* (Dec. 2019), IEEE.

[169] AKTAY, A., BAVADEKAR, S., COSSOUL, G., DAVIS, J., DESFONTAINES, D., FABRIKANT, A., GABRILOVICH, E., GADEPALLI, K., GIPSON, B., GUEVARA, M., AND OTHERS. **Google COVID-19 community mobility reports: anonymization process description (version 1.1)**. *arXiv preprint arXiv:2004.04145* (2020).

[170] ARACHCHIGE, P. C. M., BERTOK, P., KHALIL, I., LIU, D., CAMTEPE, S., AND ATIQUZZAMAN, M. **Local differential privacy for deep learning**. *IEEE Internet of Things Journal 7*, 7 (July 2020), 5827–5842.

[171] ARCOLEZI, H. H., COUCHOT, J.-F., BAALA, O., CONTET, J.-M., AL BOUNA, B., AND XIAO, X. **Mobility modeling through mobile data: generating an optimized and open dataset respecting privacy**. In *2020 International Wireless Communications and Mobile Computing (IWCMC)* (2020), pp. 1689–1694.

[172] ARCOLEZI, H. H., COUCHOT, J.-F., CERNA, S., GUYEUX, C., ROYER, G., BOUNA, B. A., AND XIAO, X. **Forecasting the number of firefighter interventions per region with local-differential-privacy-based data**. *Computers & Security 96* (Sept. 2020), 101888.

[173] BALLE, B., BARTHE, G., AND GABOARDI, M. **Privacy profiles and amplification by subsampling**. *Journal of Privacy and Confidentiality 10*, 1 (2020).

[174] BUCKEE, C. O., AND OTHERS. **Aggregated mobility data could help fight COVID-19**. *Science 368*, 6487 (Mar. 2020), 145.2–146.

[175] CARVALHO, A., CAPTIVO, M., AND MARQUES, I. **Integrating the ambulance dispatching and relocation problems to maximize system's preparedness**. *European Journal of Operational Research 283*, 3 (June 2020), 1064–1080.

[176] CERNA, S., GUYEUX, C., ARCOLEZI, H. H., COUTURIER, R., AND ROYER, G. **A comparison of LSTM and XGBoost for predicting firemen interventions**. In *Trends and Innovations in Information Systems and Technologies*. Springer International Publishing, 2020, pp. 424–434.

[177] CERNA, S., GUYEUX, C., ARCOLEZI, H. H., AND ROYER, G. **Boosting methods for predicting firemen interventions**. In *2020 11th International Conference on Information and Communication Systems (ICICS)* (Apr. 2020), IEEE.

[178] CERNA, S., GUYEUX, C., ROYER, G., CHEVALLIER, C., AND PLUMEREL, G. **Predicting fire brigades operational breakdowns: A real case study**. *Mathematics 8*, 8 (Aug. 2020), 1383.

[179] CHAMIKARA, M. A. P., BERTOK, P., KHALIL, I., LIU, D., AND CAMTEPE, S. **Privacy preserving face recognition utilizing differential privacy**. *Computers & Security 97* (Oct. 2020), 101951.

[180] DATACTIVIST. **Flux_vision**, 2020. Available online: https://datastory-datactivist.opendatasoft.com/explore/dataset/flux_vision_data-_documentation/information/ (accessed on 21 October 2021).

[181] DESFONTAINES, D. **Lowering the cost of anonymization**. PhD thesis, ETH Zurich, 2020.

[182] DUJARDIN, S., JACQUES, D., STEELE, J., AND LINARD, C. **Mobile phone data for urban climate change adaptation: Reviewing applications, opportunities and key challenges**. *Sustainability 12*, 4 (Feb. 2020), 1501.

[183] ELSALAMOUNY, E., AND PALAMIDESSI, C. **Generalized iterative bayesian update and applications to mechanisms for privacy protection**. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)* (Sept. 2020), IEEE.

[184] ERLINGSSON, Ú., FELDMAN, V., MIRONOV, I., RAGHUNATHAN, A., SONG, S., TALWAR, K., AND THAKURTA, A. **Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation**. *arXiv preprint arXiv:2001.03618* (2020).

[185] GONG, M., XIE, Y., PAN, K., FENG, K., AND QIN, A. **A survey on differentially private machine learning [review article]**. *IEEE Computational Intelligence Magazine 15*, 2 (May 2020), 49–64.

[186] GRANTZ, K. H., AND OTHERS. **The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology**. *Nature Communications 11*, 1 (Sept. 2020).

[187] HERDAĞDELEN, A., DOW, A., STATE, B., MOHASSEL, P., AND POMPE, A. **Protecting privacy in facebook mobility data during the COVID-19 response**, 2020.

[188] HOLMÉN, J., HERLITZ, J., RICKSTEN, S.-E., STRÖMSÖE, A., HAGBERG, E., AXELSSON, C., AND RAWSHANI, A. **Shortening ambulance response time increases survival in out-of-hospital cardiac arrest**. *Journal of the American Heart Association 9*, 21 (Nov. 2020).

[189] IMTIAZ, S., HORCHIDAN, S.-F., ABBAS, Z., ARSALAN, M., CHAUDHRY, H. N., AND VLASSOV, V. **Privacy preserving time-series forecasting of user health data streams**. In *2020 IEEE International Conference on Big Data (Big Data)* (Dec. 2020), IEEE.

[190] KANG, D.-Y., AND OTHERS. **Artificial intelligence algorithm to predict the need for critical care in prehospital emergency medical services**. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine 28*, 1 (Mar. 2020).

[191] KANG, Y., LIU, Y., NIU, B., TONG, X., ZHANG, L., AND WANG, W. **Input perturbation: A new paradigm between central and local differential privacy**. *arXiv preprint arXiv:2002.08570* (2020).

[192] LI, Z., WANG, T., LOPUHAÄ-ZWAKENBERG, M., LI, N., AND ŠKORIC, B. **Estimating numerical distributions under local differential privacy**. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (May 2020), ACM.

[193] LIN, A. X., HO, A. F. W., CHEONG, K. H., LI, Z., CAI, W., CHEE, M. L., NG, Y. Y., XIAO, X., AND ONG, M. E. H. **Leveraging machine learning techniques and engineering of multi-nature features for national daily regional ambulance demand prediction**. *International Journal of Environmental Research and Public Health 17*, 11 (June 2020), 4179.

[194] MERRILL, N. H., ATKINSON, S. F., MULVANEY, K. K., MAZZOTTA, M. J., AND BOUSQUIN, J. **Using data derived from cellular phone locations to estimate visitation to natural areas: An application to water recreation in new england, USA**. *PLOS ONE 15*, 4 (Apr. 2020), e0231863.

[195] NAOR, M., AND VEXLER, N. **Can Two Walk Together: Privacy Enhancing Methods and Preventing Tracking of Users**. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)* (Dagstuhl, Germany, 2020), A. Roth, Ed., vol. 156 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, pp. 4:1–4:20.

[196] NAYAK, C. **New privacy-protected facebook data for independent research on social media's impact on democracy**, 2020.

[197] OLIVER, N., AND OTHERS. **Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle**. *Science Advances 6*, 23 (Apr. 2020), eabc0764.

[198] PYRGELIS, A., TRONCOSO, C., AND CRISTOFARO, E. D. **Measuring membership privacy on aggregate location time-series**. In *Abstracts of the 2020 SIGMETRICS/Performance Joint International Conference on Measurement and Modeling of Computer Systems* (June 2020), ACM.

[199] SEZER, O. B., GUDELEK, M. U., AND OZBAYOGLU, A. M. **Financial time series forecasting with deep learning : A systematic literature review: 2005–2019**. *Applied Soft Computing 90* (May 2020), 106181.

[200] VIDAL, I. D. C., DA COSTA MENDONÇA, A. L., ROUSSEAU, F., AND MACHADO, J. D. C. **ProTECting: An application of local differential privacy for IoT at the edge in smart home scenarios**. In *Anais XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2020)* (Dec. 2020), Sociedade Brasileira de Computação.

[201] WANG, C., HORBY, P. W., HAYDEN, F. G., AND GAO, G. F. **A novel coronavirus outbreak of global health concern**. *The Lancet 395*, 10223 (Feb. 2020), 470–473.

[202] WANG, T., DING, B., XU, M., HUANG, Z., HONG, C., ZHOU, J., LI, N., AND JHA, S. **Improving utility and security of the shuffler-based differential privacy**. *Proceedings of the VLDB Endowment 13*, 13 (Sept. 2020), 3545–3558.

[203] WANG, T., LOPUHAA-ZWAKENBERG, M., LI, Z., SKORIC, B., AND LI, N. **Locally differentially private frequency estimation with consistency**. In *Proceedings 2020 Network and Distributed System Security Symposium* (2020), Internet Society.

[204] WANG, T., ZHANG, X., FENG, J., AND YANG, X. **A comprehensive survey on local differential privacy toward data statistics and analysis**. *Sensors 20*, 24 (Dec. 2020), 7030.

[205] WELLENIUS, G. A., VISPUTE, S., ESPINOSA, V., FABRIKANT, A., TSAI, T. C., HENNESSY, J., DAI, A., WILLIAMS, B., GADEPALLI, K., BOULANGER, A., AND OTHERS. **Impacts of us state-level social distancing policies on population mobility and COVID-19 case growth during the first wave of the pandemic**. *arXiv preprint arXiv:2004.10172* (2020).

[206] XIONG, X., LIU, S., LI, D., CAI, Z., AND NIU, X. **A comprehensive survey on local differential privacy**. *Security and Communication Networks 2020* (Oct. 2020), 1–29.

[207] XU, M., DING, B., WANG, T., AND ZHOU, J. **Collecting and analyzing data jointly from multiple services under local differential privacy**. *Proceedings of the VLDB Endowment 13*, 12 (Aug. 2020), 2760–2772.

[208] YANG, J., WANG, T., LI, N., CHENG, X., AND SU, S. **Answering multi-dimensional range queries under local differential privacy**. *Proc. VLDB Endow. 14*, 3 (Nov. 2020), 378–390.

[209] YANG, M., LYU, L., ZHAO, J., ZHU, T., AND LAM, K.-Y. **Local differential privacy and its applications: A comprehensive survey**. *arXiv preprint arXiv:2008.03686* (2020).

[210] YILMAZ, E., AL-RUBAIE, M., AND CHANG, J. M. **Naive bayes classification under local differential privacy**. IEEE.

[211] ZHENG, H., HU, H., AND HAN, Z. **Preserving user privacy for machine learning: Local differential privacy or federated machine learning?** *IEEE Intelligent Systems 35*, 4 (2020), 5–14.

[212] ARCOLEZI, H. H., CERNA, S., GUYEUX, C., AND COUCHOT, J.-F. **Preserving geo-indistinguishability of the emergency scene to predict ambulance response time**. *Mathematical and Computational Applications 26*, 3 (2021).

[213] ARCOLEZI, H. H., COUCHOT, J.-F., AL BOUNA, B., AND XIAO, X. **Random sampling plus fake data: Multidimensional frequency estimates with local differential privacy**. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (New York, NY, USA, 2021), CIKM '21, Association for Computing Machinery, p. 47–57.

[214] ARCOLEZI, H. H., COUCHOT, J.-F., BOUNA, B. A., AND XIAO, X. **Improving the utility of locally differentially private protocols for longitudinal and multidimensional frequency estimates**. *arXiv preprint arXiv:2111.04636* (2021).

[215] ARCOLEZI, H. H., COUCHOT, J.-F., BOUNA, B. A., AND XIAO, X. **Longitudinal collection and analysis of mobile phone data with local differential privacy**. In *Privacy and Identity Management* (Cham, 2021), M. Friedewald, S. Schiffner, and S. Krenn, Eds., Springer International Publishing, pp. 40–57.

[216] CERNA, S., ARCOLEZI, H. H., GUYEUX, C., ROYER-FEY, G., AND CHEVALLIER, C. **Machine learning-based forecasting of firemen ambulances' turnaround time in hospitals, considering the COVID-19 impact**. *Applied Soft Computing 109* (Sept. 2021), 107561.

[217] CORMODE, G., MADDOCK, S., AND MAPLE, C. **Frequency estimation under local differential privacy**. *Proceedings of the VLDB Endowment 14*, 11 (July 2021), 2046–2058.

[218] DE ALARCON, P. A., SALEVSKY, A., GHETI-KAO, D., ROSALEN, W., DUARTE, M. C., CUERVO, C., MUÑOZ, J. J., PASCUAL, J. M., SCHURIG, M., TRESS, T., DIAZ, E., DE LA CUESTA, C., AND FRIAS-MARTINEZ, E. **The contribution of telco data to fight the COVID-19 pandemic: Experience of telefonica throughout its footprint**. *Data & Policy 3* (2021).

[219] DESFONTAINES, D. **A list of real-world uses of differential privacy**, 2021. Available online: https://desfontain.es/privacy/real-world-differential-privacy.html (accessed on 02 October 2021).

[220] GARFINKEL, S. **Implementing differential privacy for the 2020 census**. USENIX Association.

[221] H. ARCOLEZI, H., CERNA, S., COUCHOT, J.-F., GUYEUX, C., AND MAKHOUL, A. **Privacy-preserving prediction of victim's mortality and their need for transportation to health facilities**. *IEEE Transactions on Industrial Informatics* (2021), 1–1.

[222] HEWAMALAGE, H., BERGMEIR, C., AND BANDARA, K. **Recurrent neural networks for time series forecasting: Current status and future directions**. *International Journal of Forecasting 37*, 1 (Jan. 2021), 388–427.

[223] KREBS, B. **T-mobile: Breach exposed ssn/dob of 40m+ people**, 2021. Available online: https://krebsonsecurity.com/2021/08/t-mobile-breach-exposed-ssn-dob-of-40m-people/ (accessed on 14 October 2021).

[224] LI, X., LIU, W., FENG, H., HUANG, K., LIU, J., REN, K., AND QIN, Z. **Privacy enhancement via dummy points in the shuffle model**. *arXiv preprint arXiv:2009.13738* (2021).

[225] LUCA, M., BARLACCHI, G., LEPRI, B., AND PAPPALARDO, L. **A survey on deep learning for human mobility**. *ACM Comput. Surv. 55*, 1 (nov 2021).

[226] MALLOUHY, R. E., GUYEUX, C., JAOUDE, C. A., AND MAKHOUL, A. **Time series forecasting for the number of firefighters interventions**. In *Advanced Information Networking and Applications*. Springer International Publishing, 2021, pp. 39–50.

[227] MARI, A. **Experian challenged over massive data leak in brazil**, 2021. Available online: https://www.zdnet.com/article/experian-challenged-over-massive-data-leak-in-brazil/ (accessed on 14 October 2021).

[228] MCCANDLESS, D., EVANS, T., QUICK, M., HOLLOWOOD, E., MILES, C., HAMPSON, D., AND GEERE, D. **World's biggest data breaches & hacks**, jan 2021. Available online: https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/ (accessed on 11 March 2021).

[229] PAPERNOT, N., AND STEINKE, T. **Hyperparameter tuning with renyi differential privacy**. *arXiv preprint arXiv:2110.03620* (2021).

[230] QU, C., KONG, W., YANG, L., ZHANG, M., BENDERSKY, M., AND NAJORK, M. **Privacy-adaptive bert for natural language understanding**. *arXiv preprint arXiv:2104.07504* (2021).

[231] RAHIMI, I., CHEN, F., AND GANDOMI, A. H. **A review on COVID-19 forecasting models**. *Neural Computing and Applications* (Feb. 2021).

[232] ROGERS, R., SUBRAMANIAM, S., PENG, S., DURFEE, D., LEE, S., KANCHA, S. K., SAHAY, S., AND AHAMMAD, P. **Linkedin's audience engagements API: A privacy preserving data analytics system at scale**. *Journal of Privacy and Confidentiality 11*, 3 (Dec. 2021).

[233] SHEN, Z., XIA, Z., AND YU, P. **PLDP: Personalized local differential privacy for multidimensional data aggregation**. *Security and Communication Networks 2021* (Jan. 2021), 1–13.

[234] STEWART, J., AND OTHERS. **Applications of machine learning to undifferentiated chest pain in the emergency department: A systematic review**. *PLOS ONE 16*, 8 (Aug. 2021), e0252612.

[235] TANG, K. J. W., ANG, C. K. E., CONSTANTINIDES, T., RAJINIKANTH, V., ACHARYA, U. R., AND CHEONG, K. H. **Artificial intelligence and machine learning in emergency medicine**. *Biocybernetics and Biomedical Engineering 41*, 1 (Jan. 2021), 156–172.

[236] TIDY, J., AND MOLLOY, D. **Twitch confirms massive data breach**, 2021. Available online: https://www.bbc.com/news/technology-58817658 (accessed on 14 October 2021).

[237] VESPE, M., IACUS, S. M., SANTAMARIA, C., SERMI, F., AND SPYRATOS, S. **On the use of data from multiple mobile network operators in europe to fight COVID-19**. *Data & Policy 3* (2021).

[238] WANG, T., LI, N., AND JHA, S. **Locally differentially private heavy hitter identification**. *IEEE Transactions on Dependable and Secure Computing 18*, 2 (Mar. 2021), 982–993.

[239] WANG, T., ZHAO, J., HU, Z., YANG, X., REN, X., AND LAM, K.-Y. **Local differential privacy for data collection and analysis**. *Neurocomputing 426* (Feb. 2021), 114–133.

[240] YANG, Z., WANG, R., WU, D., WANG, H., SONG, H., AND MA, X. **Local trajectory privacy protection in 5g enabled industrial intelligent logistics**. *IEEE Transactions on Industrial Informatics* (2021), 1–1.

[241] YOUSEFPOUR, A., SHILOV, I., SABLAYROLLES, A., TESTUGGINE, D., PRASAD, K., MALEK, M., NGUYEN, J., GHOSH, S., BHARADWAJ, A., ZHAO, J., CORMODE, G., AND MIRONOV, I. **Opacus: User-friendly differential privacy library in pytorch**. In *NeurIPS 2021 Workshop Privacy in Machine Learning* (2021).

[242] YU, D., NAIK, S., BACKURS, A., GOPI, S., INAN, H. A., KAMATH, G., KULKARNI, J., LEE, Y. T., MANOEL, A., WUTSCHITZ, L., AND OTHERS. **Differentially private fine-tuning of language models**. *arXiv preprint arXiv:2110.06500* (2021).

[243] ZHOU, X., AND TAN, J. **Local differential privacy for bayesian optimization**. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2021), vol. 35, pp. 11152–11159.

[244] BISON-FUTÉ. **Les prévisions de trafic**. Available online: https://www.bison-fute.gouv.fr (accessed on 02 February 2021).

[245] DRAKE, C. **Pyeda**. Available online: https://github.com/cjdrake/pyeda (accessed on 25 September 2019).

[246] MÉTÉO-FRANCE. **Données publiques**. Available online: https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=90&id_rubrique=32 (accessed on 02 February 2021).

[247] WORLD-HEALTH-ORGANIZATION. **WHO announces COVID-19 outbreak a pandemic**. Available online: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic (accessed on 07 September 2020).

List of Figures

List of Tables

Document generated with LaTeX and:
the LaTeX style for PhD Thesis created by S. Galland — http://www.multiagent.fr/ThesisStyle
the tex-upmethodology package suite — http://www.arakhne.org/tex-upmethodology/

école doctorale **sciences pour l'ingénieur et microtechniques**

Title: Production of Categorical Data Verifying Differential Privacy: Conception and Applications to Machine Learning

Keywords: Differential privacy, Local differential privacy, Categorical data, Machine learning.

Abstract:

Private and public organizations regularly collect and analyze digitalized data about their associates, volunteers, clients, etc. However, because most personal data are sensitive, there is a key challenge in designing privacy-preserving systems to comply with data privacy laws, e.g., the General Data Protection Regulation. To tackle privacy concerns, research communities have proposed different methods to preserve privacy, with Differential privacy (DP) standing out as a formal definition that allows quantifying the privacy-utility trade-off. Besides, with the local DP (LDP) model, users can sanitize their data locally before transmitting it to the server.
The objective of this thesis is thus two-fold: $\mathbf{O}_1$) To improve the utility and privacy in multiple frequency estimates under LDP guarantees, which is fundamental to *statistical learning*. And $\mathbf{O}_2$) To assess the privacy-utility trade-off of machine learning (ML) models trained over differentially private data.

Titre : Production of Categorical Data Verifying Differential Privacy: Conception and Applications to Machine Learning

Mots-clés : Confidentialité différentielle, Confidentialité différentielle locale, Données catégorielles, Apprentissage automatique.

Résumé :

Les organisations privées et publiques collectent et analysent régulièrement des données numérisées sur leurs associés, volontaires, clients, etc. Cependant, comme la plupart des données personnelles sont sensibles, la conception de systèmes préservant la vie privée pour se conformer aux lois sur la confidentialité des données, par exemple le règlement général sur la protection des données, constitue un défi important. Pour résoudre les problèmes de confidentialité, les communautés de chercheurs ont proposé différentes méthodes de préservation de la confidentialité, la confidentialité différentielle (DP) se distinguant comme une définition formelle qui permet de quantifier le compromis entre confidentialité et utilité. En outre, avec le modèle de confidentialité différentielle locale (LDP), les utilisateurs peuvent sanitisé leurs données localement avant de les transmettre au serveur.
L'objectif de cette thèse est donc double : $\mathbf{O}_1$) Améliorer l'utilité et la confidentialité des estimations de fréquences multiples sous garanties LDP, ce qui est fondamental pour *l'apprentissage statistique*. Et $\mathbf{O}_2$) Évaluer le compromis vie privée-utilité des modèles d'apprentissage machine (ML) entraînés sur des données différentiellement privées.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France