UNIVERSITY OF CRETE
DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF SCIENCES AND ENGINEERING

Decentralized privacy-preserving data sharing: The case of wearable data.

by

Thomas Marchioro

Ph.D. Dissertation

Presented in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

March 2023

UNIVERSITY OF CRETE

DEPARTMENT OF COMPUTER SCIENCE

Decentralized privacy-preserving data sharing: The case of wearable data.

Ph.D. Dissertation Presented

by **Thomas Marchioro**

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

APPROVED BY:

Author: Thomas Marchioro

Supervisor: Evangelos P. Markatos, Professor, University of Crete

---

Committee Member: George Pallis, Professor, University of Cyprus

---

Committee Member: Šarūnas Girdzijauskas, Professor, KTH Royal Institute of Technology

---

Committee Member: Maria Papadopouli, Professor, University of Crete

---

Committee Member: Athena Vakali, Professor, Aristotle University of Thessaloniki

---

Committee Member: Yannis Stylianou, Professor, University of Crete

---

Committee Member: Polyvios Pratikakis, Professor, University of Crete

---

Department Chairman: Antonis Argyros, Professor, University of Crete

March 2023

Acknowledgments

I would like to take this opportunity to express my gratitude and appreciation to all those who have supported me throughout these three years of PhD.

First and foremost, I would like to thank my supervisor Evangelos, for his valuable guidance throughout this process. His expertise and feedback have been instrumental in shaping my research and, hopefully, making it successful.

I am also grateful to my thesis committee members George and Sarunas, for their feedback and suggestions that have helped me to refine my work and improve its quality.

I would like to extend my sincere gratitude to the RAIS project that gave me the opportunity to collaborate with some of the smartest and most capable colleagues I have ever worked with: Andrei, Lodovico, Sofia, and George. I want to thank Andrei in particular for his invaluable help and contribution to this work. Without his assistance, this thesis would not have been possible.

Last but not least, I would like to express my appreciation to my family and friends. To my family, for their constant support and encouragement, even while I was miles away. To my old friends, in particular Stefano, Alessandro, Fabio, Marco, Enrico, and Gabriella, who have been with me since before my PhD and never forgot about me. To my new friends, whom I met during my PhD program, in particular Michalis, Areti, Georgia, Katerina, and Eva, for making my time in Greece more enjoyable and meaningful.

Thank you all for being a part of my thesis journey. Your support has been priceless and deeply appreciated.

Abstract

Commercial wearable devices such as fitness trackers and smartwatches have gained momentum as powerful tools for health research. However, collecting and publishing such data raises privacy concerns that must be addressed. In this thesis, we investigate the privacy risks associated with collecting and publishing wearable data for health-related applications. Specifically, we examine the potential for wearable data to be used as a "fingerprint" to de-anonymize people who produce them. We mainly focus on the possibility of re-identifying fitness tracker users that have participated in health-related experiments such as observational studies. To mitigate these threats, we propose various defense mechanisms, including centralized and decentralized anonymization approaches, to protect individuals in wearable datasets. Although both approaches are studied, the thesis primarily focuses on the latter, as they enable personalized data sanitization processes that can meet different privacy requirements. Moreover, the proposed decentralized solutions provide theoretical privacy guarantees based on the concept of differential privacy. In this work, we demonstrate the feasibility of decentralized mechanisms for anonymizing wearable data in health studies and promote them as a practical tool for collecting and publishing such data.

Keywords: **data privacy**, **decentralized solutions**, **wearable data**

Supervisor Evangelos Markatos, Professor, University of Crete

Advisory committee Evangelos Markatos, Professor, University of Crete;
George Pallis, Professor, University of Cyprus;
Šarūnas Girdzijauskas, Professor, KTH Royal Institute of Technology

Περίληψη

Εμπορικές φορετές συσκευές όπως φιτνεσς τρασκερς και σμαρτωατσηες έχουν συγκεντρώσει την προσοχή ως δυνατά εργαλεία για έρευνες στον τομέα της υγείας. Όμως, η συλλογή και δημοσίευση τέτοιου τύπου δεδομένων εγείρει ζητήματα σε θέματα διασφάλισης προσωπικών δεδομένων τα οποία πρέπει να αντιμετωπιστούν. Σε αυτή την διατριβή, ερευνούμε κινδύνους σε θέματα διασφάλισης προσωπικών δεδομένων που προέρχονται από την συλλογή και δημοσίευση δεδομένων από εφαρμογές υγείας φορετών συσκευών. Συγκεκριμένα, εξετάζουμε το ενδεχόμενο δεδομένα προερχόμενα από φορετές συσκευές να χρησιμοποιηθούν ως ένα 'δακτυλικό αποτύπωμα' που μπορεί να καταργήσει την ανωνυμία των ανθρώπων που τα παρήγαγαν. Επικεντρονώμαστε κυρίως στην πιθανότητα επαναταυτοποίησης χρηστών εφαρμογών τύπου φιτνεσς τρασκινγ οι οποίοι έχουν λάβει μέρος σε πειράματα υγείας, όπως σε μελέτες παρατήρησης. Προκειμένου να μετριαστούν τέτοιου είδους απειλές, παραθέτουμε διάφορους μηχανισμούς άμυνας, συμπεριλαμβανομένων ζεντραλιζεδ και δεζεντραλιζεδ προσεγγίσεων, με σκοπό την προστασία ατόμων ευρισκόμενων σε δατασετς φορετών συσκευών. Παρότι και οι δύο προσεγγίσεις μελετώνται, η διατριβή ως επί το πλείστων εστιάζει στην δεύτερη, καθώς αυτή επιτρέπει διαδικασίες σανιτιζατιον των προσωπικών δεδομένων, οι οποίες έχουν την δυνατότητα να μπορούν να πληρούν διάφορες προϋποθέσεις θεμάτων ασφάλειας. Επιπλέον, οι προτεινόμενες αποκεντρωτικές προσεγγίσεις παρέχουν εγγυήσεις ασφάλειας και σε θεωρητικό επίπεδο μέσω της έννοιας της διφφερεντιαλ πριασψ. Σε αυτή την εργασία, δείχνουμε την δυνατότητα εφαρμογής αποκεντρωτικών μηχανισμών προς ανωνυμοποίηση δεδομένων προερχόμενων από μελέτες υγείας μέσω φορετών συσκευών και τους προτείνουμε ως ένα πρακτικό εργαλείο για την συλλογή και δημοσιοποίηση τέτοιας φύσης δεδομένων.

Contents

List of Figures

List of Tables

Publications

The main body of this thesis includes the following publications:

I Marchioro T, Kazlouski A, Markatos E. *User Identification from Time Series of Fitness Data*. In SECRYPT 2021 (pp. 806-811). Scitepress.

II Kazlouski A, Marchioro T, Markatos E. *What your Fitbit Says about You: De-anonymizing Users in Lifelogging Datasets*. In 19th International Conference on Security and Cryptography (SECRYPT), July 11-13, 2022, Lisbon, Portugal 2022 (pp. 806-811). Scitepress.

III Marchioro T, Kazlouski A, Markatos E. *How to Publish Wearables' Data: Practical Guidelines to Protect User Privacy*. Studies in Health Technology and Informatics. 2022 May 1;294:949-50.

IV Marchioro T, Giaretta L, Markatos E, Girdzijauskas Š. *Federated Naive Bayes under Differential Privacy*. In 19th International Conference on Security and Cryptography (SECRYPT), July 11-13, 2022, Lisbon, Portugal 2022 (pp. 170-180). Scitepress.

V Yfantidou S, Karagianni C, Efstathiou S, Vakali A, Palotti J, Giakatos DP, Marchioro T, Kazlouski A, Ferrari E, Girdzijauskas Š. *LifeSnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild*. Scientific Data. 2022 Oct 31;9(1):1-9.

VI Marchioro T, Kazlouski A, Markatos E. *Practical Crowdsourcing of Wearable IoT Data with Local Differential Privacy*. Submitted to 8th IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI) 2023.

This work also produced other related publications, which were not included in the main body of the thesis:

VII Kazlouski A, Marchioro T, Manifavas H, Markatos E. *Do you know who is talking to your wearable smartband?*. Integrated Citizen Centered Digital Health and Social Care. 2020 Jul:142.

VIII Kazlouski A, Marchioro T, Manifavas H, Markatos E. *I still See You! Inferring Fitness Data from Encrypted Traffic of Wearables*. In HEALTHINF 2021 Feb (pp. 369-376).

IX Kazlouski A, Marchioro T, Manifavas H, Markatos E. *Do partner apps offer the same level of privacy protection? The case of wearable applications*. In 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) 2021 Mar 22 (pp. 648-653). IEEE.

X Giaretta L, Savvidis I, Marchioro T, Girdzijauskas Š, Pallis G, Dikaiakos MD, Markatos E. *PDS^2: A user-centered decentralized marketplace for privacy preserving data processing*. In 2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW) 2021 Apr 19 (pp. 92-99). IEEE.

XI Kazlouski A, Marchioro T, Markatos E. *I just wanted to track my steps! Blocking unwanted traffic of Fitbit devices*. Accepted at 12th International Conference on the Internet of Things (IoT '22).

XII Giaretta L, Marchioro T, Markatos E, Girdzijauskas Š. *Towards a decentralized infrastructure for data marketplaces: narrowing the gap between academia and industry*. In Proceedings of the 1st International Workshop on Data Economy 2022 Dec 9 (pp. 49-56).

Chapter 1
Introduction

Commercial wearable fitness trackers, such as smartbands and smart watches, have been gaining popularity as a solution to monitor daily activity and promote physical exercise [57]. Having real-time feedback on their activity, along with the possibility of sharing such information with peers, has shown to be beneficial for people, who feel motivated to exercise and compete with each other [19]. Moreover, consumer-level fitness trackers have proven their usefulness in health research, offering a non-invasive and cost-effective way to monitor study participants [46]. They are utilized, for example, to assess the effectiveness of physical activity interventions [32, 88, 108] and to analyze the progresses of patients during rehabilitation [99, 116].

However, while gathering and sharing of activity data presents numerous advantages for health research, this has also raised several concerns regarding the privacy aspect of data collection [20]. Fitness trackers record a variety of sensitive personal parameters from their users, including physical activity level, heart rate measurements, and sleeping time [31, 56, 100]. These records are known in literature as *wearable data*. Not only these data allow to monitor lifestyle and habits of individual users, but, depending on how frequently they are sampled, they may also reveal a person's activity on a specific day and time. Additionally, some of the measured parameters may contain traces of identifying information: for instance, most commercial fitness trackers estimate burned calories based on the user's gender, age, height, and weight [33, 115, 120]. In essence, wearable fitness trackers produce both identifying and sensitive data. Considering also that users are generally not aware of the privacy implications of sharing their wearable data [2, 23], this makes them the perfect target for attacks aimed at breaching privacy.

With these premises in mind, this thesis analyzes both privacy threats to consumer-level fitness trackers and potential defense strategies, focusing mainly on applications in fitness and health research. In particular, we[1] examine attacks that are are aimed

[1]Throughout the manuscript, I use "we" to underline the contribution of collaborators and supervisors to this research. However, the content of the thesis reflects my views only.

at re-identifying device users based on their wearable data. More specifically, we consider the possibility of using data collected by wearables as a "fingerprint" to identify anonymous users in public datasets.

The defense mechanisms investigated in this work, instead, aim to guarantee the anonymity of the individuals included in a dataset of wearable data. In this manuscript, the terms *anonymization*, *de-identification*, and *sanitization* will be used interchangeably to denote procedures that aim to separate the identity of an individual from his or her disclosed data. Furthermore, we distinguish between *centralized* and *decentralized* anonymization mechanisms. In this thesis, we consider "centralized" a mechanism in which anonymity is enforced by a central authority (e.g., the organization who conducts the study that sanitizes the data after collection). On the other hand, we call "decentralized" a mechanism in which the de-identification process is carried by the data owner (e.g., users sanitizing their own data before submitting them to the organizer).

Arguably, decentralized mechanisms are generally preferable in health studies, as they do not require participants to entrust their personal data to the study organizer or to a third party. Furthermore, decentralization puts the device users in charge of the anonymization process, enabling a more personalized data sanitization and allowing to meet different privacy requirements. Therefore, while we study both centralized and decentralized solutions to guarantee anonymity, we mainly focus on the latter ones and on their applicability in health studies based on wearables.

1.1 Research questions

In summary, the research questions that we investigate in this thesis are as follows:

> **RQ1**: What are the practical risks of participating in a health study that makes use of wearables? To what extent do these risks impact the privacy of a user?
>
> **RQ2**: What can data collectors/publishers do to protect wearable data before disclosing them?
>
> **RQ3**: What can device users do in order to protect their data before disclosing them? Are there viable decentralized solutions to anonymize/sanitize wearable data while preserving their utility for health researchers?

1.2 Thesis contributions

The main contributions of this thesis in addressing each research question can be summarized as follows:

Research Question	Contribution	Publications
RQ1	Re-identification attacks against wearable data	Publications I and II
RQ2	Guidelines to publish wearable data	Publications III and V
RQ3	Decentralized algorithms for differentially private data collection	Publications IV and VI

Table 1.1: List of contributions of the thesis with related publications.

- Investigating RQ1, we explore possible approaches that an attacker may use to uncover a user's identity by analyzing activity records (e.g., steps taken, calories burned, and distance traveled). Specifically, we study the effectiveness of re-identification attacks that leverage such records to predict a user's identity or personal details. To our knowledge, we are the first to analyze practical attacks and evaluate them on real datasets of wearable data..

- For RQ2, we formulate a set of guidelines aimed at helping wearable data collectors protect sensitive information before disclosure. These guidelines are primarily based on general principles and do not provide theoretical guarantees. On the other hand, they can easily be utilized by practitioners who lack technical expertise in the privacy domain, and can thus limit the risk of private information being exposed.

- Finally, to address RQ3, we develop novel paradigms for privacy-preserving data sharing based on well-established decentralized techniques. These paradigms are specifically designed for two applications: (i) computing aggregated statistics, such as average values, cumulative distribution, and p-values, and (ii) training specific classes of machine algorithms, such as Naive Bayes classifiers. In particular, we propose for a crowdsourcing platform design that collects data randomized with differential privacy guarantees. We demonstrate that this platform enables the estimation of the target aggregated statistics while preserving privacy. Additionally, we have devised a federated version of the Naive Bayes algorithm. Although these paradigms cover a limited number of applications, they allow for the extraction of useful information from wearable data in relevant use cases while simultaneously providing theoretical privacy guarantees.

The findings presented in this thesis have been previously published in peer-reviewed conferences. Table 1.1 provides a schematic summary that indicates the publications associated with each contribution of the thesis.

The results reported in the next chapters are the outcome of a joint work with fellow researchers, and part of a larger project titled *Real time analytics for the Internet*

of Sports (RAIS). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813162. However, the content of this thesis reflects our view only. The European Commission/ Research Executive Agency are not responsible for any use that may be made of the information it contains.

1.3 Novelty of the contributions

Most of the contributions of this thesis are based on prior works and well-established privacy solutions. For this reason, we explicitly mention our original contribution to the state of the art in each chapter.

- Previous works suggested that a large collection of wearable data could constitute a fingerprint for device users [20, 123]. However, our research represents a pioneering effort in devising actual attacks that target wearables. These were inspired by previous attacks that exploited similarity between records, but exploited specific properties of fitness trackers data. Specifically, both our attacks capitalize on the how estimated calorie consumption and distance traveled relate to personal user information, such as gender, height, and weight.

- The guidelines we propose for publishers are primarily based on well-established privacy principles, such as data minimization and k-anonymity. Nevertheless, our findings reveal that public datasets of wearable records often fail to adhere to these principles. This discovery has prompted us to emphasize the prevailing misconceptions and pitfalls made by publishers, aiming to raise awareness about the risks associated with uncontrolled disclosure of such data.

- We introduce a novel design for a crowdsourcing platform aimed at collecting wearable data while preserving local differential privacy. While the randomization algorithm we employ to enforce differential privacy is well-established, our original contribution lies in the development of the data collection pipeline. This aspect is particularly crucial for enhancing the privacy-utility tradeoff, which represents the main challenge faced by differential privacy-based solutions.

- We develop a federated version of the Naive Bayes algorithm, a commonly used machine learning technique. Although the federated learning paradigm is predominantly employed in neural networks, our approach distinguishes itself by utilizing different local queries and aggregation algorithm. This framework enhances both the training efficiency, and improves the privacy-utility tradeoff when differential privacy is applied.

1.4 Thesis structure

The rest of this dissertation is organized as follows. Chapter 2 covers fundamental concepts that are used in the main body of the thesis. These include applications of wearable data in health research, such as comparative studies and machine learning, and well-established anonymization techniques, such as k-anonymity [121] and differential privacy [27]. Chapter 3 describes novel attack vectors that we designed to compromise the privacy of wearable data. These attacks demonstrate that standard anonymization techniques are insufficient to protect privacy in datasets of wearable records. In chapter 4 we present a set of guidelines that data publishers can use to mitigate privacy threats against wearable data. These guidelines are practical solutions that can be adopted in a centralized setting. Chapter 5 focuses on decentralized solutions based on differential privacy that device users can adopt to protect their data. The chapter explores two primary applications of wearable data in health research, namely comparative studies and machine learning, and demonstrates how the proposed solutions can securely collect wearable information while preserving its utility. Finally, chapter 6 provides conclusions drawn from this thesis.

Chapter 2
Background

This chapter covers fundamental concepts that are necessary to understand the rest of the dissertation. In the first part, we discuss the most prominent applications of wearable data to health research. We focus on comparative studies, which are used to assess strategies for physical activity interventions and patient rehabilitation, and machine learning applications such as stress and injury prediction. Furthermore, we describe public datasets of wearable records that are used in the next chapters. The second part instead, motivates and describes well-established anonymization techniques such as k-anonymity and differential privacy.

2.1 Notation

This thesis explores attacks and defense mechanisms, which are founded on fundamental statistical concepts. Therefore, we need to introduce a minimal amount of mathematical notation to facilitate comprehension and minimize ambiguity. Some key symbols will maintain the same meaning throughout the entire dissertation, while others will be defined within each paragraph according to the context.

Time series data The primary focus of this research is on *time series of wearable records*, which are indexed by multiple time indices denoted as $t = 1, \dots, T$. Unless otherwise specified, we assume a sampling rate of one record per day. A time series of wearable records is a sequence x of T records. Depending on the application at hand, each record may contain different parameters such as steps, calories, distance, heart rate, etc. Generally, x can be treated as a collection of $T \times d$ values, where d represents the number of parameters per record. While parameters like steps, calories, and distance are typically integers, we will treat them as real numbers and denote x as belonging to $\mathbb{R}^{T \times d}$. The record at time t in the time series x is denoted as $x[t] \in \mathbb{R}^d$. A specific parameter, or feature, f of $x[t]$ is denoted as $x_f[t]$. When considering multiple time series $x_1, \dots, x_n$, the feature index $f = 1, \dots, F$ follows the data index $i = 1, \dots, n$.

In certain cases, however, this notation may become cumbersome, particularly when writing equations that pertain to fixed features and time indices. For the sake of clarity, in such instances we may opt to use x instead of $x_f[t]$, explicitly stating this simplification in the text, by writing $x \in \mathbb{R}$.

Datasets Another fundamental concepts of this thesis is *datasets*. These are denoted with the letter D and are simply collections of data with common structure. In most cases, we will work with datasets of time series data, which comprise time series by n distinct users, i.e. $D = (x_i, \ldots, x_n)$. The space of possible datasets is $\mathcal{D}$.

Random variables Part of the privacy preservation approaches we adopt rely on randomization mechanisms. A random mechanism $\mathcal{A}$ refers to a stochastic function that maps either a dataset $D \in \mathcal{D}$ or a value $x \in \mathbb{R}$ to a random value in $\mathbb{R}$. To analyze the statistical properties of this value, we treat it as a *random variable*. Random variables can assume values within continuous or discrete ranges, following a probability distribution. The specific value taken by the random variable is called an *observation*. In our notation, we represent random variables using capital letters X, Y, or Z, and their observations with the corresponding lower case letters x, y, and z.

2.2 Wearable data and health research

Wearables allow researchers to collect data on individuals' daily activity levels and other health-related behaviors. These data can be used to study a range of health outcomes, including chronic disease management, physical activity, and sleep patterns. Researchers can study wearable data to gain insights into the relationship between certain behavioral patterns and health outcomes.

As already mentioned in the introduction, fitness trackers have been applied to a variety of applications in health research. While certain wearable devices are designed for specific purposes, such as measuring oxygen levels in the blood or providing accelerometer data, this work predominantly centers on consumer-level wearables and their applications in research. This focus is motivated by the belief that data generated from consumer-level wearables can have a transformative impact on research: millions of these devices are sold every year, and each device user can opt to donate his or her data to health research. Having access to such a large and diverse pool of data would imply a high reliability for the studies that are based on such data.

Consumer-level fitness trackers are mainly utilized in two applications. The first one is observational studies [110], which employ the devices to track groups of people and detect changes in their behaviors. These studies can be useful for evaluating the effectiveness of an intervention or therapy. The second application is the development

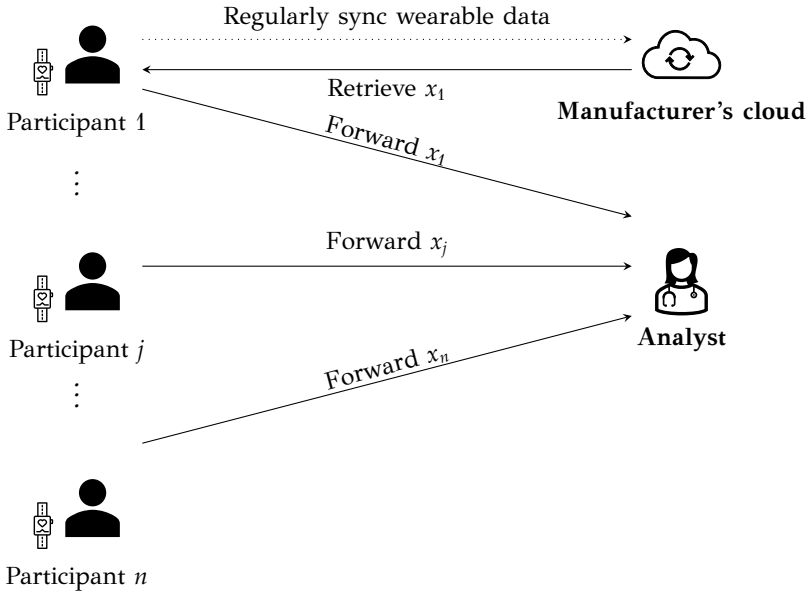


Figure 2.1: Typical data collection pipeline for wearable data in scientific studies. The data of each user are collected and logged by the fitness tracker, and are then uploaded to the manufacturer's cloud. To create a dataset, the analyst requests that each study participant retrieve their data from the cloud and forward it to the analyst. These are merged into the collection $(x_1, \ldots, x_n)$, which is the final dataset.

of machine learning models, which use wearable data to identify stress or illnesses [96] and predict symptoms [42].

In these studies, users of wearable fitness trackers simply retrieve their data from the cloud of the device manufacturer, and forward them to the analyst, as shown in figure 2.1. The analyst collects all the user submissions and merges them into a single dataset.

2.2.1 Comparative studies

Observational studies can have different scope and methodology. In this thesis, we mainly consider comparative studies, in which researchers compare the characteristics, behaviors, or outcomes of two or more groups of individuals. Comparative studies can be used to assess the effectiveness of physical activity interventions by having one group undergoing the intervention and another group used as control [32, 46, 108].

In these studies, wearables are used to provide objective measurements about the two groups of participants. The main parameters that are recorded are the number

of steps taken per day, energy consumption, distance covered, and active minutes, which are recorded by most commercial fitness trackers [133]. Occasionally, also sleep cycles [87] and heart rate [22,92] are monitored, but typically these measurements are considered less reliable when taken by commercially available devices.

In many cases, studies that rely on consumer-level fitness trackers aim to assess the effectiveness of patient rehabilitation strategies. A research by Kelly et al., for example, [65] compares different call schedules and their effects on the sleep and physical activity of the surgical residents. This work measures the number of steps taken by the patients to quantify physical activity and compares it to their sleeping cycles. Other studies [101,102] investigated the effect of physical activity interventions on cancer survivors by conducting randomized trials. Wearables were used to measure steps and active minutes for participants in both the intervention and control groups. These are just few examples of how these devices can be beneficial to health studies, and many others can be found in the related literature [43,63,84,113].

Statistical significance and t-test Comparative studies not only measure differences between groups of individuals, but also establish whether such differences are statistically significant. This is done by running statistical hypothesis tests, in which two possible hypotheses are considered:

- H_0 (also, called the "null hypothesis"): the two groups are not significantly different, and any variation in measurement between the two groups happened by chance;
- H_1: the two groups are significantly different.

Neither of these hypotheses can be rejected with certainty. However, it is possible to estimate the probability of H_0 being true. More specifically, one can compute an upper bound to the probability of H_0, which is known in literature as the p-value. If the p-value is below a certain threshold, hypothesis H_1 is accepted as true, meaning that the difference between the two groups is statistically significant.

The most popular statistical test that is employed in comparative studies is the Student's t-test [67]. This test is based on the assumption that the measurements follow a normal (i.e., Gaussian) distribution, which is reasonable in most applications. The normal distribution naturally arises in many measurements of physical and biological phenomena [14,129], and also characterizes physical parameters such as height [6] as well as the duration of physical exercise [3]. Additionally, even if the normality assumption is not met, results on t-test still generalize well on other distributions [12]. Under the assumption of normally-distributed data, embracing the null hypothesis means that the data belong to a single Gaussian distribution with a given mean $\mu \in \mathbb{R}$

and standard deviation $\sigma \in \mathbb{R}$. Conversely, rejecting the null hypothesis implies that the measurements of the two groups follow two different Gaussians with different means μ_1, μ_2 and standard deviations σ_1, σ_2.

The p-value is determined by the total number n of measured data points and by the value of the t-statistic, which is computed as

$$t_{\text{stat}} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}\sqrt{\frac{2}{n}}}, \tag{2.1}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the averages of the two groups and $\hat{\sigma}$ is the empirical standard deviation of the whole sample.

2.2.2 Machine learning

Wearable data also play an important role in the development of machine learning models for healthcare [8, 100, 111]. Machine learning can be used as a data-driven approach to "train" predictive models that leverage wearable data to forecast health outcomes. For example, some research has proposed using wearables to predict injuries in athletes due to fatigue or stress based on motion data [136]. Other works have focused on detecting stress by combining activity data with smartphone usage records [96]. Other research targeted more specific applications, e.g., aiming to predict flares in arthritic patients based on decreases in physical activity levels [42]. Recent studies have also suggested the possibility of using wearables to predict COVID-19 waves [5]. Finally, some works [95,115] use machine learning approaches to correct measurement errors made by wearables and improve the quality of the data produced by these devices.

In opposition to current trends, many works that train machine learning models with wearable data do not use neural network architectures. Instead they adopt simpler models such as linear regression [15], Naive Bayes [96,137], and decision trees [48,98]. There are two main reasons behind this choice. The first one is that simpler model typically provide more robust results even when the available data are scarce. This is often the case for applications that rely on wearable, since the dataset needs to be collected by the institution conducting the study. The other reason is that these models are more explainable. Naive Bayes treats separate features[1] independently, allowing to determine the contribution of each individual feature to the final prediction. Decision trees can be interpreted as a set of binary questions that are asked about the input data. Having this kind of interpretation allows to infer general rules based on the

[1] in the context of machine learning, the term "features" is used to denote the input variables used to make a prediction.

trained model. Neural networks, on the other hand, are more complex and are mostly used as black-box models. This makes them suitable for applications such as image captioning or text generation, for which obtaining an accurate or credible output is more important than understanding the intuition behind these predictions.

In this thesis, we rely mainly on two types of machine learning models. The first type is Naive Bayes models, which are widely used to make predictions based on wearable data. In chapter 5, we design a privacy-preserving version of the Naive Bayes algorithm that can be used to train models without requiring centralized access to individual data points. The other type of model that we use is neural networks. As previously mentioned, neural networks are often not suitable for wearable-based predictions, since they lack robustness and interpretability. However, they proved to be a powerful tool for user de-anonymization, as shown in chapter 3.

Naive Bayes A Naive Bayes [45] model is a parametric classification model, which is based on the assumption that the features – or attributes – of data points are independent. The class prediction is based on the maximum a posteriori probability criterion, meaning that the predicted class $\hat{y}$, chosen among a number of possible classes $1, \ldots, C$, should maximize the posterior probability $p_{Y|X}(y|x)$. In other words, $\hat{y}$ should be the class to which the data point $x \in \mathbb{R}^F$ is "most likely" to belong. This can be done, applying the Bayes theorem, by maximizing the product of the prior probability p_Y with the conditional probabilities $p_{X_f|Y}(x_f|y)$ for all features $f = 1, \ldots, F$ i.e.,

$$\hat{y} = \arg \max_{y=1,\ldots,C} p_Y(y) \prod_{f=1}^{F} p_{X_f|Y}(x_f|y). \tag{2.2}$$

During training, all these probabilities are estimated based on the distribution of the training data. Prior probabilities $p_Y(y)$ are computed for each feature as

$$p_Y(y) = \frac{n_y}{\sum_{y'} n_{y'}} \tag{2.3}$$

where n_y is the number of data points in class y. For conditional probabilities, Naive Bayes employs a different training and prediction processes, distinguishing between categorical features (the value of which can belong to a finite number of categories, like gender) and numerical features (which may in principle take any value, like height). Conditional probabilities for categorical features are estimated similarly to priors, using the counts of each category $m_{x_f y}$ to compute

$$p_{X_f|Y}(x_f|y) = \frac{m_{x_f y}}{n_y}. \tag{2.4}$$

The most common approach for numerical features, instead, is to assume that they follow a normal distribution [17,93], which is characterized by a mean μ_{fy} and standard deviation σ_{fy} that is different for every class. These are estimated from the data as

$$\mu_{fy} = \frac{1}{n_y} \sum_{x \text{ in class } y} x_f \tag{2.5}$$

and

$$\sigma_{fy} = \sqrt{\frac{1}{n_y - 1} \sum_{x \text{ in class } y} (x_f - \mu_{fy})^2}. \tag{2.6}$$

The conditional probability is proportional to the probability density function of the normal distribution with parameters μ_{fy} and σ_{fy}, i.e.,

$$p_{X_f|Y}(x_f|y) \propto \frac{1}{\sqrt{2\pi}\sigma_{fy}} \exp\left(-\frac{(x_f - \mu_{fy})^2}{2\sigma_{fy}^2}\right). \tag{2.7}$$

Neural networks It is not an exaggeration to say that neural networks have become the most important tool in modern machine learning research. The vast majority of advancement in the field of artificial intelligence are based upon neural network architectures. While entire books can be written on these models, in this paragraph we focus on describing their basic functioning. A neural network model is characterized by three main pillars [40], namely the architecture, the loss function, and the optimization algorithm:

- *Architecture*: Neural networks are complex models but are essentially the composition of basic building blocks, which are called "layers". Layers are mostly parametric linear operations followed by a fixed non-linear operation, called activation function. Training a neural network model implies finding the optimal parameters that characterize the linear operations in each layer. Intuitively, more layers imply a higher representation capability for the model. The way in which layers are combined defines the architecture of the model.

- *Loss function*: The parameters of a neural network model are tuned to minimize a certain objective, called loss function. This function represents a penalty for the mistakes made by the model. Minimizing it essentially means reducing the amount of mistakes that the model makes in its prediction task. A fundamental requirement for this function is that it should be differentiable with respect to the model's output. Another typical requirement is that the loss function should be convex.

- *Optimization algorithm*: While several optimization algorithms are used to train neural networks, these are all variants of a basic algorithm called stochastic gradient descent (SGD). SGD is based on the idea of gradually reaching the optimal loss value by updating the parameters towards the opposite direction to the gradient of the loss. Since the gradient represents the direction of growth for a function, following the opposite direction implies reaching points were the loss function decreases, eventually reaching the minimum value if the function is convex. When descending the gradient, however, the loss is not computed on the entire dataset, but on a random batch of data points. This is why the procedure is stochastic.

Two main types of layers that we use in our work are fully-connected (also known as dense or linear layers) and LSTM layers [119]. Fully-connected layers are the simplest type of layer and simply consist in a matrix multiplication with the input and the addition of a bias term. LSTM layers still perform linear operations but are designed to process time series of data. They utilize a state vector to store temporal information while processing the input sequence. For the loss function, we use the cross-entropy, which penalizes mistakes in classification tasks. Finally, we use Adam [68] for the optimization algorithm, which guarantees a more stable training by accounting for two types of momentum when computing the descending direction.

2.2.3 Public datasets of wearable records

Most works, including both comparative studied and machine learning research, do not publish the wearable data that they collected. There are many reasons to not publish private datasets. In the case of wearable data, one of them is a lack of guidelines on how to publish them. Due to the scarce availability of datasets, the scope of our work is limited to the few datasets that we could retrieve from public sources. A typical structure of these datasets is reported in figure 2.2, comprising one table with the participants' details and another one with their activity records. We utilized three publicly available datasets and one that was collected by the RAIS consortium, LifeSnaps. The characteristics of these datasets are detailed below.

Fitbit Connections [94] This dataset was collected by the Open Humans Foundation[2], and counts 40 users who shared their data for a period ranging from 17 to 3509 days. The records are aggregated at a daily level, i.e., hourly details for the activity data is not available. Personal information about the participants includes weight and height. Furthermore, some of the participants reported their first name – or even full name – which uniquely identified them. In our experiments we used the reported

[2] http://openhumansfoundation.org

names to determine the gender of the participants in the cases where this was not ambiguous.

Furberg et al. [34] This dataset was collected via the Amazon Mechanical Turk crowd-sourcing platform, and comprises data from 35 Fitbit users. As the dataset does not have an official name, we refer to it using the names of its publishers. Personal attributes include height and weight of the participants, while their gender is not reported. The number of recorded days per participant ranges from 2 to 49.

PMData [122] This dataset was created through a lifelogging experiment that lasted for five months and involved 16 athletes who used the Fitbit Versa 2 wristband. The dataset includes gender, age, height, and weight information for all participants except one, for whom only weight information is missing. Unlike the other two datasets, this dataset was generated during a more controlled experiment, resulting in few missing records for each participant, with the exception of one athlete who sustained an injury during the early stages of the experiment. The number of days recorded per participant in this dataset ranges from 80 to 152. Measurements are available at the default sampling rate recorded by Fitbit. Steps are stored as hourly records, while calories are displayed in a minute-by-minute format.

LifeSnaps This is a multimodal dataset of Fitbit records collected by the RAIS consortium. It comprises 71 participants recruited from 4 different countries, namely Cyprus, Greece, Italy, and Sweden. The dataset contains information about gender and body mass index (BMI) for each participant. Furthermore, aggregated information about age group, height, and education level has been made available. For each user, the dataset stores about two months (≈60 days) worth of activity records, aggregated hourly.

2.3 Data privacy

In principle, privacy may seem a straightforward concept. However, delving deeper into the topic, one will quickly realize that people have different ideas of what privacy is. Some believe that having their pictures uploaded online is a violation of their privacy, while some others do not see any problem with that. Dissimilar notions of privacy exist not only in our everyday lives, but also in academic research. Every year, hundreds of papers get published that propose "privacy-preserving" or "privacy-enhancing" methods to protect data. Yet, most of these papers follow their own definition of privacy, which may be completely unrelated to the others.

#	Gender	Age
001	Female	24
002	Male	28
003	Male	24
004	Female	22
005	Female	26
006	Male	26
007	Male	26
008	Female	27
009	Male	24
010	Male	24
011	Male	27

Date	Record	001	002	...	011
22/05/03	steps	17873	9243	...	14306
	distance	14424	6136	...	10343
	calories	4007	1999	...	3703
	sleep	8	7	...	8
22/05/04	steps	13118	10246	...	13235
	distance	10584	7109	...	9646
	calories	3529	2095	...	3381
	sleep	10	4	...	8
...	...	...	...	...	...
22/07/02	steps	14312	11489	...	9037
	distance	11460	7631	...	6546
	calories	3747	2223	...	3324
	sleep	8	7	...	9

Figure 2.2: Example of dataset of wearable records. Typically such datasets contain both demographic information (left) and time series of activity information (right). A user ID is assigned to each participant so that their demographic information can be associated with their records.

In this thesis, we focus on a specific aspect of privacy, which is *anonymity*. Anonymity essentially consists in keeping someone's identity hidden, which implies that any kind of "fingerprint" needs to be concealed. This is a particularly relevant aspect in the context of health studies, as they typically target participants with sensitive conditions. If someone has taken part in a research study on diabetic patients, it may indicate that this person is affected by diabetes. For this reason, our work focuses on collecting and disclosing wearable data and getting the most out of them while maintaining people's identity protected.

In the rest of this section, we define the concept of anonymous data and the requirements that these data must satisfy. Additionally, we describe well-known attacks that aim to break individual's anonymity, as well as standard defense mechanisms that mitigate such attacks. We mainly focus on protection mechanisms that provide theoretical guarantees, such as k-anonymity and differential privacy. Furthermore, henceforth we use the terms "anonymity" and "privacy" interchangeably to make the text more readable and less repetitive.

Anonymous data We refer to *anonymous data* as any type of data that include personal information but cannot be uniquely linked to a specific individual. But why do we even need anonymous data? If the goal is to protect personal information, the most

Name	Email	Gender	Age	Diabetes
Alice	alice@mail.com	Female	24	Yes
Bob	bob@mail.com	Male	28	Yes
Charlie	charlie@mail.com	Male	24	No
Diane	diane@mail.com	Female	22	Yes
Ester	ester@mail.com	Female	26	No
Fred	fred@mail.com	Male	26	No
George	george@mail.com	Male	26	Yes
Hannah	hannah@mail.com	Female	27	Yes
Ian	ian@mail.com	Male	24	No
John	john@mail.com	Male	24	No
Kenneth	kenneth@mail.com	Male	27	Yes

Table 2.1: Example of survey data collected during and experiment. Typically participants are required to provide contacts and demographic information. In this table, names, email addresses, and age are direct identifiers, as they contain unique entries for one or more participants.

effective approach is to refrain from disclosing such information altogether.

The reason why anonymous data are necessary is that personal information can be really useful, especially in the medical domain. Therefore, an additional requirement that anonymized information should satisfy is that the data should maintain some utility. Often the usability of anonymous data and their privacy are contrasting goals, and a good privacy mechanism should reach a reasonable compromise between the two. This is known in literature as the *privacy-utility tradeoff*.

2.3.1 Linking attacks and k-anonymity

A first naive approach to protect personal data may be to just remove direct identifiers from a dataset and publish the rest of the information as is. Direct identifiers are attributes[3] that directly identify a specific individual. Examples of direct identifiers include names, social security numbers, email addresses, and phone numbers. These attributes can be used to link a specific individual to their personal information and are considered sensitive. For example, suppose that we have conducted a study on a group of participants to compare exercise habits of young diabetic and non-diabetic people, and that the collected data are stored in a spreadsheet as reported in table 2.1. Clearly, names and email addresses are direct identifiers and must be removed.

[3]One can think of "attributes" as columns of a database.

Attacker's knowledge about her targets.

Target	Gender	Age
Alice	Female	24
Bob	Male	28
Fred	Male	26
George	Male	26

#	Gender	Age	Diabetes
001	Female	24	Yes
002	Male	28	Yes
003	Male	24	No
004	Female	22	Yes
005	Female	26	No
006	Male	26	No
007	Male	26	Yes
008	Female	27	Yes
009	Male	24	No
010	Male	24	No
011	Male	27	Yes

Table 2.2: Typical threat model for a linking attack: the attacker knows that certain targets of hers are present in the database. She utilizes the quasi-identifiers in her possession to compare the two tables and eventually re-identify the targets. In this example, the attacker is able to re-identify Alice and Bob, but not Fred and George.

Quasi-identifiers However, it is well known that removing direct identifiers is not sufficient to protect anonymity. An attacker may exploit another type of personal information, called quasi-identifiers, that allows to re-identify a target member of a dataset. Quasi-identifiers are attributes that do not directly identify an individual on their own, but may be used in combination with other attributes to re-identify individuals. Examples of quasi-identifiers include gender, birth dates, and zip codes. An attacker who possesses such information about her target can possibly re-identify him. This type of threat, in which an attacker is aware that her target is present in a dataset and re-identifies him based on additional information in her possession, is called a linking attack. The typical threat model for linking attacks consists in the attacker having an additional table that connects quasi-identifiers with the identity of the targets.

In the example reported in table 2.2, the attacker performs a linking attack using gender and age as quasi-identifiers. Notice that she is able to re-identify Alice and Bob, but not Fred and George. The reason is that Fred and George share the same tuple of quasi-identifiers. This concept of identical quasi-identifiers can be leveraged to protect the members of a dataset from re-identification.

***k*-anonymity** *k*-anonymity [121] is a privacy model that aims to protect the identities of individuals in a dataset by ensuring that they cannot be re-identified based on quasi-identifiers. A formal definition of *k*-anonymity can be formulated as follows.

Consider a dataset D in which the attributes $f_1, \ldots, f_m$ are quasi-identifiers. A dataset achieves k-anonymity if for all $x \in D$, there are $x', \ldots, x^{(k-1)} \in D$ such that

$$(x_{f_1}, \ldots, x_{f_m}) = (x'_{f_1}, \ldots, x'_{f_m}) = (x^{(k-1)}_{f_1}, \ldots, x^{(k-1)}_{f_m}). \tag{2.8}$$

For example, a dataset of individuals with gender and age as quasi-identifiers achieves 3-anonymity if each member of the dataset shares the same gender and age with at least two other individuals.

In the k-anonymity model, the dataset is trusted to a "curator", who is in charge of protecting the identity of its members. It is assumed that the curator is able to determine which attributes of the dataset can be considered quasi-identifiers. The curator groups the dataset members based on their quasi-identifiers and counts how many individuals are present in each group. These groups are anonymized by replacing the original quasi-identifiers with generalization and suppression until each group contains at least k members. Suppression simply means removing an attribute, whereas generalization means that the attribute is replaced with less specific values (e.g., the birth date "25 Dec 1995" could be generalized to just "Dec 1995"). Both these operations remove information from the dataset while increasing the level of privacy.

In k-anonymity, the value of k is a parameter that regulates the privacy-utility tradeoff for the dataset. Specifically, if a dataset satisfies k-anonymity with respect to its quasi-identifiers, an attacker has at best $1/k$ probability of re-identifying her target. A large value of k means that individuals are more protected but also requires to sacrifice more useful information. For small datasets, most of the information is typically removed to achieve just 2- or 3-anonymity.

In the example from table 2.1, 2-anonymity can be enforced by dividing members in two age groups, 20–24 and 25–29. In order to achieve a higher value of k, instead, either the gender or the age would need to be completely suppressed, as no further generalization can be applied. One may think to further generalize the the age column into a single interval 20–29. However, this interval would not provide any information about the dataset entries, so it does not need to be stored as a column.

On the other hand, applying k-anonymity to datasets with many participants and few attributes can protect the anonymity of the participants with a small utility cost. Due to its simplicity and to the efficient of algorithms that enforce it, k-anonymity is widely adopted and constitutes one of the pillars for data privacy [24].

#	Gender	Age	Diabetes
001	Female	20–24	Yes
002	Male	25–29	Yes
003	Male	20–24	No
004	Female	20–24	Yes
005	Female	25–29	No
006	Male	25–29	No
007	Male	25–29	Yes
008	Female	25–29	Yes
009	Male	20–24	No
010	Male	20–24	No
011	Male	25–29	Yes

QID	#
(Female, 20–24)	2
(Female, 25–29)	2
(Male, 20–24)	3
(Male, 25–29)	4

Table 2.3: Example of table achieving 2-anonymity (left) with corresponding counts of quasi-identifier tuples (right). In order to achieve 3-anonymity, either the gender or the age need to be completely suppressed.

2.3.2 Privacy leaks and differential privacy

Despite its popularity and effectiveness, k-anonymity is not a universal solution. A first limitation of using this model is its proneness to various types of inference attacks. Considering the example in table 2.3, an attacker would be able to successfully carry out a homogeneity attack [105] on the dataset. This type of attack exploits homogeneity in the sensitive information for a specific anonymous group. By observing that all the individuals with quasi-identifiers (Female, 20–24) have diabetes, an attacker could determine that her 24-year old female target, Alice, also has diabetes. In other words, even if the anonymity of individual members is preserved, one may still be able to glean sensitive information. Another strategy that an attacker may adopt is considering multiple participants. For instance, since the attacker is targeting 3 out of 4 members in the (Male, 25–29) group, and in that group 3 participants have diabetes, she can conclude that at least two of her targets have diabetes.

Another weakness of the k-anonymity model is the assumption that the curator is able to determine which attributes can be considered quasi-identifiers. In the example presented above, quasi-identifiers were clearly personal attributes like gender and age. However, it has been shown by different works that people can be re-identified based on their movie preferences [89,90] or on the topology of their social network connections [91], which at a first glance may not seem information that needs to be protected. In general, identifying quasi-identifiers is an open problem and any attribute could be considered a quasi-identifier in principle. However, applying k-anonymity for all the

attributes in a dataset often renders the data unusable. This is especially true for datasets of time series records, for which each timestamp is a potential quasi-identifier.

Membership inference and privacy We just mentioned how privacy leaks may happen in different ways depending on the strategy adopted by the attacker and on what she is trying to infer. However, there is a general criterion to determine if a dataset protects against individual-level privacy leaks. This was proposed by Cynthia Dwork and colleagues [27] when introducing the concept of differential privacy, which is explained below. The main idea is that privacy leaks occur when an individual is present in a dataset, and its presence affects the content of the dataset. In other words, a member of a dataset is subject to privacy risks only when its presence affects the dataset.

Differential privacy Differential privacy [29] can be a confusing concept at first, as it refers both to a mathematical definition of privacy and to a set of techniques for implementing that definition. However, the core idea to keep in mind is that to apply differential privacy typically means applying a controlled amount of noise to the data. More specifically, differential privacy mostly consists in randomizing the output of an algorithm that runs on a private dataset, so that the presence of a specific individual in the dataset cannot be inferred.

The rigorous definition of this concept is based on the notion of "neighboring" or "adjacent" datasets. Two datasets D and D' are adjacent if they differ by exactly one row. A randomized algorithm $\mathcal{A}$ is differentially private if for all adjacent datasets D, D', the following inequality

$$\Pr[\mathcal{A}(D) \in O] \leq e^{\varepsilon} \Pr[\mathcal{A}(D') \in O]. \tag{2.9}$$

holds for all the subsets O of the output range $\text{Range}(\mathcal{A})$ of $\mathcal{A}$. The output range is the set of possible values that can be returned by the randomization algorithm and may vary depending on its design. Intuitively, an algorithm is differentially private (or, quantitatively, ε-differentially private) if replacing one row of the dataset[4] does not change "too much" the output distribution of the algorithm. The parameter $\varepsilon > 0$ is called "privacy budget" and quantifies the limit to this change.

It is important to notice that since the definition holds for *any* pair of adjacent datasets, D and D' can be swapped in equation 2.9, leading to both an upper and a lower bound for the distribution of $\mathcal{A}(D)$:

$$e^{-\varepsilon} \Pr[\mathcal{A}(D') \in O] \leq \Pr[\mathcal{A}(D) \in O] \leq e^{\varepsilon} \Pr[\mathcal{A}(D') \in O]. \tag{2.10}$$

[4]Typically, one row corresponds to one individual.

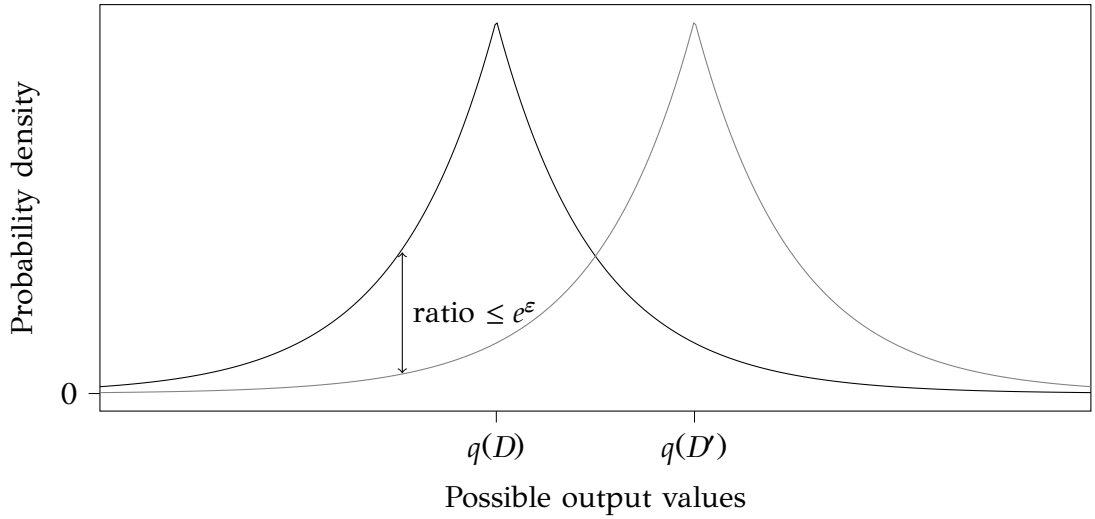


Figure 2.3: Visual representation of the differential privacy definition for a given value of ε. The two curves represent the probability density of a randomization algorithm for a query q computed on adjacent datasets D and D'. The ratio between the two curves can never exceed e^{ε}.

A visual representation of these bounds is shown in figure 2.3, which depicts the Laplace mechanism, described later in this section. The algorithm computes a query (i.e., a function $q : \mathcal{D} \to \text{Range}(q) \subset \mathbb{R}$) and randomizes it so that the final output achieves ε-differential privacy. The mechanism maps $q(D)$ to a random value according to a continuous probability distribution. This probability has its peak at the actual value of $q(D)$ and gradually decreases for values that are far from $q(D)$. The output range of this mechanism is $(-\infty, \infty)$.

The figure also shows that, applying the Laplace mechanism to two adjacent datasets D and D', we get two distinct distributions. However, the ratio between their probability density functions for each point in the range never exceeds e^{ε}.

Local differential privacy The formulation of differential privacy provided above applies to datasets of multiple data points. However, in applications where data points need to be anonymized locally, the notion of local differential privacy (LDP) is often preferred. This is equivalent to the notion of differential privacy with one single row. In this case for any pair of data points x, x', the output of the mechanism should satisfy

$$\Pr[\mathcal{A}(x) \in O] \leq e^{\varepsilon} \Pr[\mathcal{A}(x') \in O], \ \forall O \subseteq \text{Range}(\mathcal{A}). \tag{2.11}$$

Properties Two fundamental properties [86] utilized throughout the discussion are as follows:

- **P1**: If m independent $\frac{\varepsilon}{m}$-differentially private algorithms are run on a same dataset D, then any function of their output satisfies ε-differential privacy.

- **P2**: If ε-differentially private algorithms are computed on disjoint subsets of a dataset $D^{(1)}, \ldots, D^{(N)}, D^{(i)} \cap D^{(j)} = \emptyset, i \neq j$, then any function of their outputs provides ε-DP.

Next, we define two well-known mechanisms that achieve differential privacy, applicable to scalar queries or values, denoted as $q(D) \in \mathbb{R}$ or $x \in \mathbb{R}$. By leveraging properties P1 and P2, these mechanisms can be used for multi-valued queries and data. This can be done by simply applying the mechanisms independently to each scalar component and appropriately distributing the privacy budget.

Laplace mechanism A well-known mechanism to enforce differential privacy is the Laplace mechanism [28]. The Laplace mechanism L_ε for a query q computed on a dataset D consists in simply adding Laplace noise to the query with scale inversely proportional to ε:

$$\mathcal{A}(D) = L_\varepsilon(q(D)) = q(D) + Z, \quad Z \sim \text{Lap}(0, \frac{\Delta}{\varepsilon}). \tag{2.12}$$

The value Δ is the maximum variation of the query between adjacent datasets, also called the *sensitivity* of the query. The Laplace mechanism can also be used to achieve LDP by simply replacing $q(D)$ with the value $x \in \mathbb{R}$ in equation 2.12. The output of the Laplace mechanism can take any values in $(-\infty, +\infty)$. However, values that are farther from the original value of $q(D)$ (or x) are reached with lower probability.

Piecewise mechanism In local differential privacy, another popular mechanism is the piecewise mechanism. The Piecewise mechanism was originally introduced by [128] and improved by [140]. The core idea of the mechanism is to randomize a scalar input $x \in [-1, 1]$ to a limited range $[-A, A]$, $A \in \mathbb{R}_+$, according to a piecewise-uniform probability density function (PDF). The PDF is divided into a high-density region $(L(x), R(x))$ which is constructed around x, and a low density region $[-A, L(x)] \cup [R(x), A]$, which covers the rest of the range $[-A, A]$. More formally, the mechanism acts as follows:

$$\mathcal{A}(x) = \text{PW}_\varepsilon(x) = Y, Y \sim p(y|x) \tag{2.13}$$

where $p(y|x)$ is a PDF defined as

$$p(y|x) = \frac{\tau(e^{\varepsilon}-1)}{2(\tau+e^{\varepsilon})^2}\begin{cases} e^{\varepsilon}, & \text{if } y \in (L(x), R(x)) \\ 1, & \text{if } y \in [-A, L(x)] \cup [R(x), A] \\ 0, & \text{otherwise} \end{cases} \tag{2.14}$$

where

$$A = \frac{(e^{\varepsilon}+\tau)(\tau+1)}{\tau(e^{\varepsilon}-1)},\ L(x) = \frac{(e^{\varepsilon}+\tau)(x\tau-1)}{\tau(e^{\varepsilon}-1)},\ R(x) = \frac{(e^{\varepsilon}+\tau)(x\tau+1)}{\tau(e^{\varepsilon}-1)}. \tag{2.15}$$

Normally, τ is suitably chosen depending on the values of x and ε. However, in our experiments we adopt the sub-optimal solution $\tau = e^{\varepsilon/3}$ [140]. Although the mechanism is defined for an input in $[-1, 1]$, it can be trivially applied to any input in a bounded range $[x_{\min}, x_{\max}]$. In essence, the original sample is scaled to $[-1, 1]$, calculated according to eq. 2.14-2.15, and then rescaled back. Since the Piecewise mechanism outputs values in $[-A, A]$, the rescaled output falls in the range $[x_{\min} + x_{\max}\frac{1-A}{2}, x_{\min} + x_{\max}\frac{1+A}{2}]$.

Chapter 3
Attacks on wearable data

In this chapter, we focus on the privacy risks associated with wearable data and explore the extent to which individuals can be re-identified from public datasets of this data. We first outline the personal and sensitive information that may be contained in these datasets, highlighting the potential implications of this information falling into the wrong hands.

Then, in order to demonstrate the feasibility of attacks against public datasets of wearable data, we devise two re-identification attacks that leverage specific background knowledge about a target individual. The first attack involves using additional records of wearable data collected from the target victim to compare with records in the dataset, with the aim of identifying which dataset participant is the target victim. The second attack involves inferring certain physical characteristics, such as gender, height, and weight from the wearable data and using this information to re-identify the target.

The outcomes of these attacks provide evidence of the vulnerability of wearable data to privacy violations and underscore the importance of implementing appropriate safeguards to protect individuals' personal and sensitive information.

3.1 Personal information contained in wearable data

As outlined in the introduction, wearable data collected by fitness trackers carry a wealth of personal information, some of which may be highly sensitive. These devices collect behavioral data such as the number of steps taken, calories burned, and exercise habits. These data may reveal patterns regarding individual's habits and routine. Furthermore, most commercial fitness trackers are also able to monitor health parameters, such as heart rate, sleep patterns, and in some cases stress levels measured by electrodermal activity (EDA). Combining all these parameters together and collecting them at a high sampling rate yields a quite accurate overview of the fitness status and lifestyle of the device user. While this could be positive in certain situations where patients need to be monitored for prolonged periods in an unobtrusive manners, disclosing so

much information inevitably raises privacy concerns [20].

Most works primarily focus on the private companies that are collecting such data, since the standard behavior of most commercial fitness trackers is to upload the collected data on the manufacturer's cloud [59–61]. In this thesis, on the other hand, we focus on data collected by research institution and universities. Many research works, especially in the medical and sports fields, relied on this devices to conduct behavioral research, e.g., monitoring patients and athletes. The great majority of such studies only reported the main findings, but few of them instead made the collected data publicly available, which are those reported in chapter 2. In the latter case, the datasets were released in different formats and the type and granularity of information varies depending on the application. However, they all carried two main types of information:

- **demographic information**, containing details about the participants' background, such as gender, age, height, weight;
- **activity and health information**, typically stored as time series of steps taken, calories burned, distance covered, hours slept, which ranged over multiple days.

In our work we thoroughly studied possible unwanted leaks of information that may derive from careless disclosure of such datasets.

3.1.1 Identifying information

A first critical issue that we observed is that most of the publicly available datasets do not even apply k-anonymity to demographic information. Attributes such as gender, age, and height of the dataset participants can surely be considered quasi-identifiers. If these are not properly anonymized through generalization and suppression, the identity of the participants might get exposed. However, even in the case where participants are properly clustered into anonymous groups based on their demographic data, we concluded that their anonymity is still not guaranteed. As mentioned in 2.3.2, a main limitation of k-anonymity is that it assumes that quasi-identifiers can be easily recognized by the data curator. In reality, it is hard to determine what information is "identifying" for a dataset participant, as this strongly depends on the additional information held by the attacker. Suppose, for example, that the attacker knows that Bob ran a marathon on May 4 and took over 50000 steps. If there is only one member of the dataset with over 50000 steps on May 4, the attacker can easily tell that it is Bob.

This is not the only way in which the attacker can identify an anonymous participant. Another source of information is the relationship between steps taken and

calories burned. In commercial fitness trackers, the number of steps is directly calculated using accelerometer measurements. Calorie consumption, on the other hand, is estimated using a combination of steps, activity information (e.g., heart rate), and personal information inserted by the device user, such as gender, age, height, and weight. In a similar way, the covered distance is estimated by combining steps and height information. As a consequence, records of such information may be used to trace an individual's personal characteristics.

Harris-Benedict calorie estimation The relationship between estimated calories and personal information is even more evident when considering basal metabolic rate (BMR), i.e., the number of calories burned by a person at rest to maintain basic bodily functions for one day. Preliminary experiments that we conducted showed that the BMR estimated by Fitbit closely follows the Harris-Benedict equation [44]. The Harris-Benedict equation, developed in 1918 by American scientists James Arthur Harris and Francis Gano Benedict, estimates BMR as a linear combination of age, height, and weight, i.e.

$$\text{BMR} = \alpha \cdot \text{age} + \beta \cdot \text{height} + \gamma \cdot \text{weight} + \gamma. \tag{3.1}$$

The coefficients of this formula vary between males and females, meaning that also gender is taken into account. We used PMData, the only available dataset that contains all the required parameters, to validate Harris-Benedict formula on Fitbit records. In order to extract BMR from Fitbit records, we examined the minute-by-minute calories recorded during sleep and kept the most frequent value. This was multiplied by $24 \times 60 = 1440$ to obtain the total number of basal calories consumed in one day, i.e., the BMR. Figure 3.1 shows a comparison between the BMR extracted from Fitbit records and the estimated value using the Harris-Benedict equation. The difference between the two never exceeds the 200 kcal threshold, suggesting that calories indeed relate to personal information, and can thus be used to infer it.

3.1.2 Related work

The vast majority of the studies that investigated sensitive inference from activity information mainly rely on raw sensor data such as accelerometer and gyroscope [25, 64, 97, 114].

Some papers explored the possibility of learning sensitive information from wearable records of steps taken, calories burned, and distance covered. Most of these papers, however, consider the standard k-anonymity threat model, in which individuals can be identified only using personal attributes such as gender, height, and weight [4, 76]. Torre et al. [123] explored the correlation between these measurements and discussed the need to consider them in the anonymization of wearable data. To our knowledge,

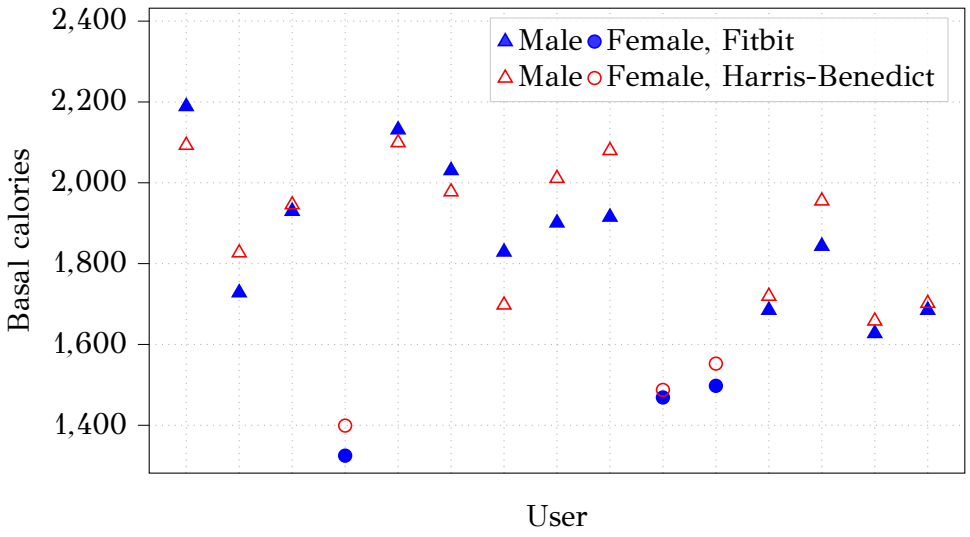


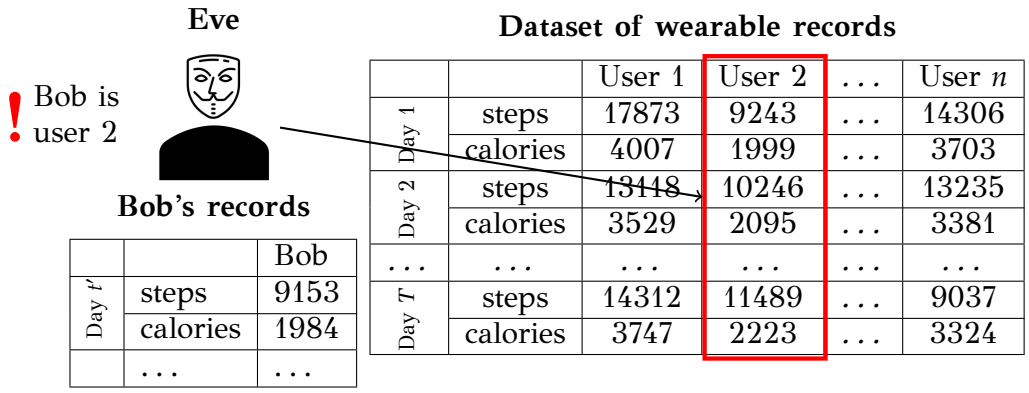
Figure 3.1: Comparison of basal metabolic rate (BMR) estimated using Harris-Benedict equations (empty marked points) and extracted by Fitbit data (full marked points) for the participants of PMData.

no one designed and evaluated realistic re-identification attacks against datasets of wearable records.

3.2 Re-identification attacks on wearable data

We have already mentioned that if an attacker knows her target's activities on a particular day, she can easily re-identify him if that day is included in the dataset's time series. This alone is a realistic threat, but in our research we bring the attack vector one step further. We posit a scenario in which the attacker knows that a target of hers is in the dataset, which ranges over a certain period of time. However, we assume that the attacker does not possess activity information on her target within that same temporal scope. Another assumption is that the dataset's demographic attributes were completely removed, so the attacker cannot uncover the participants' identity based on them. In this scenario, we explore two types of re-identification strategies [62, 81]:

- **Re-identification via record linking**: In this case, we assume that the attacker possesses activity information on her target. However, the activity information gathered by the attacker and the time series in the dataset range over different periods in time.
- **Re-identification via inference of demographic information**: In this case, we assume that the attacker has some background knowledge regarding the tar-



		Bob
Day t'	steps	9153
	calories	1984
	...	...

		User 1	User 2	...	User n
Day 1	steps	17873	9243	...	14306
	calories	4007	1999	...	3703
Day 2	steps	13118	10246	...	13235
	calories	3529	2095	...	3381
...	...	...	...	...	...
Day T	steps	14312	11489	...	9037
	calories	3747	2223	...	3324

Figure 3.2: Threat model for re-identification based on record linking. The attacker (Eve) knows that her target (Bob) is present in the public dataset. She has access to additional records of her target that she has collected from other sources (e.g., social media or fitness communities), and she compares them with each user in the dataset.

get, namely gender, height, and weight. She leverages the activity information contained in the dataset to deduce these characteristics for each participant and de-anonymizes the target based on them.

In both cases we assume that the attacker has access to a public dataset of wearable records collected by n individuals. We denote with $x_i \in \mathbb{R}^{T \times d}$ the time series of samples collected by the i-th user, and the overall dataset can be described as the ordered collection $D = (x_1, x_2, \ldots, x_n)$. The entry of x_i at day t, denoted with $x_i[t]$, is a collection of F activity parameters, e.g., steps taken and calories burned.

3.2.1 Re-identification via record linking

In the case of record linking, depicted in figure 3.2, the adversary tries to infer which time series between $x_1, \ldots, x_n$ is more similar to another time series x' that she has collected from her target victim. This additional time series x' may have been obtained due to a link between the attacker and the victim, or gathered by victim's posts on social media or online fitness communities [2]. If the time range of x' and the time series in the public dataset overlap, re-identifying the target is trivial. Thus, we focus on the case where there is no time overlapping between the records in D and x'.

This attack is designed to de-anonymize participants based on daily records. The default template provided by Fitbit and other fitness apps on their social networks

is aggregated by day, thus it is more likely for an adversary to obtain this kind of information. However, data collected for scientific purposes may have been sampled at a higher frequency, e.g., every hour or even minute by minute. In such cases, a malicious actor can carry out the attack described below by first aggregating the samples at a daily level. This way, the attacker would obtain precise details about the habits and lifestyle of the victim.

Methodology In order to link the data of the target user x' to his corresponding time series in the aggregated database, we employ a minimum distance approach. We compute a normalized euclidean distance between each record in the dataset $x_i[t]$, $i = 1, \ldots, n$, $t = 1, \ldots, T$, and each record $x'[t']$, $t' = 1, \ldots, T'$, of the target's time series:

$$d(x_i[t], x'[t']) = \sqrt{\sum_{f=1}^{F} \frac{1}{\sigma_f}(x_{i,f}[t] - x'_f[t'])^2}. \tag{3.2}$$

While computing this distance metric, all parameters (steps, calories, distance etc.) are normalized by their respective standard deviation σ_f, so that their contribution is equal. Otherwise, parameters with higher values, such as steps, would have a bigger impact on the calculated distance.

For each timestamp t' of the target's time series x', we match it to the closest record $x_i[t]$ in the dataset (for any value of t) and this implies that the i-th participant could be the attacker's target. The final guess is made using a majority rule, i.e., choosing the participant with most matched records

$$\hat{\imath} = \arg \max_{i \in \{1,\ldots,n\}} \left| \{t' : \arg\min_j d(x_j[t], x'[t']) = i\} \right|. \tag{3.3}$$

The attack is successful if $x_{\hat{\imath}}$ is actually the time series of the target victim.

Experimental evaluation In order to determine the effectiveness of the attack, we evaluate its success rate, i.e., the probability to correctly re-identify the target in the dataset. We estimate the success rate through a Monte Carlo simulation of the attack for a varying number of participants n, testing the approach on the Furberg et al. and PMData datasets. For both datasets, we use the first half (time-wise) of the records to simulate the dataset D, while we extract the target's records x' from the second half. Participants with an excessively small number of records were removed, since the purpose is to test the effectiveness of the attack when such information is available. We evaluate the attack performance focusing on a scenario in which only steps and calories are available (i.e., $F = 2$), making the threat model applicable to most existing

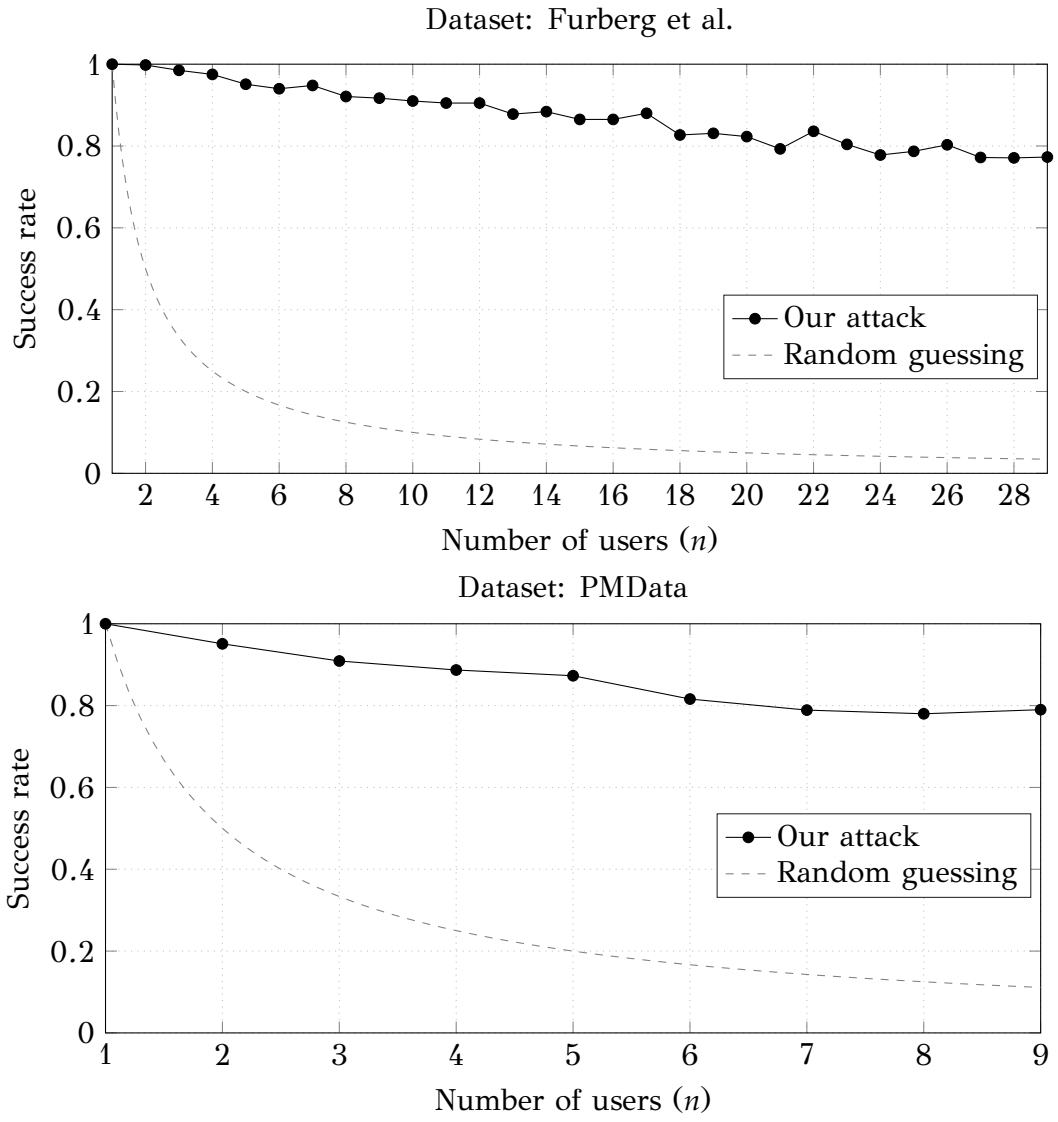


Figure 3.3: The success rate of our linking attack based on steps and calories records, estimated for varying number n of users. We ran a Monte Carlo simulation with 10,000 trials for each value of n. Parameters used to link two records: steps and calories.

datasets of wearable records.

The Monte Carlo simulation consists in repeating the following steps for a fixed number of trials n_{trials}:

1. Sample n participants $i_1, \ldots, i_n$ uniformly at random from the dataset, and among them choose one (also uniformly at random) to be the target i'.

2. Select the time series $x_{i_1}, \ldots, x_{i_n}$ and x' accordingly.
3. Compute all the distances between the records in $x_{i_1}, \ldots, x_{i_n}$ and the records in x'.
4. Link x' to one of the participants among $i_1, \ldots, i_n$ according to 3.3.
5. Verify if the guess is correct (successful trial).

The success rate of the attack is estimated as

$$\text{Success rate} = \frac{\text{Number of successful trials}}{\text{Total number of trials}}. \tag{3.4}$$

This metric can be interpreted as the level of "accuracy" or effectiveness that the attacker achieves in de-anonymizing their target.

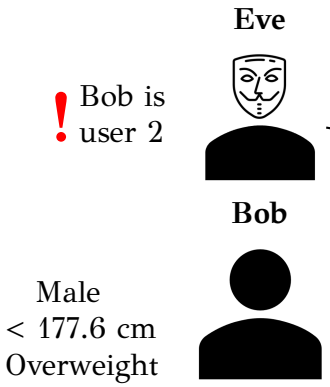
The results of our experiments for the Furberg et al. and PMData datasets are reported in figure 3.3. For each value of n, we ran a Monte Carlo simulation with $n_{\text{trials}} = 10,000$ trials. In Furberg et al., the attack turned out to be remarkably effective: even when the target victim was selected at random among 28 participants, re-identification was successful for more than 78% of the trials, while a random guess would yield only $1/n = 3.4\%$ success rate. PMData showed slightly lower results, possibly due to the lower diversity of the participants. Being a dataset of athletes, the participants tend to have more similar activity habits and body types. Nonetheless, the success rate for $n = 9$ participants is still above 78%, way above the $1/n = 11.1\%$ of a random guessing strategy.

3.2.2 Re-identification via inference of demographic information

In this case, depicted in figure 3.4, the adversary predicts demographic attributes of each participant of the dataset based on the sequences $x_1, \ldots, x_n$, and utilizes these attributes to de-anonymize participants with unique characteristics. To be more precise, the adversary utilizes these records to answer three binary questions, i.e.,

- q_{gender}, whether a participant is male or female (only two options provided by Fitbit);
- q_{bmi}, whether a participant is overweight (BMI ≥ 25) or not.
- q_{height}, whether a participant's height is above or below the European average of 177.6 cm;

In our threat model we assume that all the demographic attributes have been completely removed. However, if the adversary is able to infer the answer to the above



Dataset of wearable records

		User 1	User 2	...	User n
Day 1	steps	17873	9243	...	14306
	distance	14424	6136	...	10343
	calories	4007	1999	...	3703
Day 2	steps	13118	10246	...	13235
	distance	10584	7109	...	9646
	calories	3529	2095	...	3381
...	...	...	...	...	...
Day T	steps	14312	11489	...	9037
	distance	11460	7631	...	6546
	calories	3747	2223	...	3324

Figure 3.4: Threat model for re-identification based on inference of demographic attributes. The attacker (Eve) knows that her target (Bob) is present in the dataset. She leverages the wearable records to infer information regarding the gender, height and weight of each participant and she compares it with background information that she knows about Bob.

binary questions, she might be able to use them to single out her victim, making pointless the anonymization of such attributes.

Methodology In order to predict the answer to the binary questions we adopted an approach based on machine learning. More specifically, we train a different neural network model for each binary question. These models receive individual records as input and output a binary answer $\hat{b} \in \{0, 1\}$ for each question. Also in this case, we take into account multiple records by adopting a majority rule. Given the predictions $\hat{b}_i[t] = q(x_i[t])$, $t = 1, \ldots, T$, on T samples of the i-th participant, the final answer $\hat{r}_i$ predicted for that participant is

$$\hat{r}_i = \begin{cases} 1, & \text{if } \frac{1}{T}\sum_{t=1}^{T} \hat{b}_i[t] > \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

Experimental evaluation We first evaluated how accurately an attacker can predict the answer to the binary questions q_{gender}, q_{bmi}, q_{height}. We used the Fitbit Connections and Furberg et al. datasets as training data, and tested the resulting models on PMData. We adopted the same multi-layer perceptron architecture for all models, which consists in 3 layers with 120, 60 and 2 output neurons, respectively. We used the ReLU

Question	Record-wise accuracy	User-wise accuracy
q_{gender}	0.925	1.000 (16/16)
q_{bmi}	0.731	1.000 (15/15)
q_{height}	0.821	0.938 (15/16)

Table 3.1: Accuracy of the trained models in answering the binary questions q_{gender}, q_{bmi}, q_{height}. Record-wise accuracy represents the fraction of records for which the prediction is successful, while user-wise accuracy shows the overall successful predictions for the participants after applying the majority rule. Our models successfully predicted the gender and BMI questions for all participants, while one short person (below 177.6 cm) was predicted as tall.

activation function between each pair of intermediate (hidden) layers, and applied the Softmax activation to the output layer. All models were trained using the cross-entropy loss, batch size 64 and Adam optimizer with learning rate 10^{-3}. Furthermore, the best model was chosen by performing a 5-fold cross-validation, dividing the training data with an 80/20 split. The accuracy results are reported in table 3.1. We distinguish between record-wise accuracy, which is the accuracy obtained on individual records, and user-wise accuracy, which is the overall accuracy obtained when making the final prediction for a participant using the majority rule. Remarkably, both q_{gender} and q_{bmi} were correctly predicted for all the participants in PMData. For q_{height}, one short participants (i.e., below the threshold of 177.6 cm) was predicted as tall. Nevertheless, the results suggest that these three physical characteristics can be reliably predicted by our models. This implies that an attacker may employ them to de-anonymize participants. Being three binary questions, answering them can reveal the identity of at most $2^3 = 8$ individuals. Still, this represents a risk for participants who have more distinct characteristics. For example, this attack may be used to find a short overweight participant in a dataset where most participants are tall and not overweight.

In the case PMData, if all the questions are correctly predicted, the dataset can be divided into 6 groups, as shown in table 3.2. Among these, 3 groups contain a single individual, meaning that 3 individuals are uniquely identifiable by their answers to $(q_{\text{gender}}, q_{\text{bmi}}, q_{\text{height}})$. Essentially, once the characteristics are correctly predicted, the de-anonymization approach is the same one discussed in the case of k-anonymity. Using k-anonymity terminology, this is a perfect example of datasets whose quasi-identifiers are not obvious. Even if a curator decides to suppress gender, height, and weight of the participants, an attacker can partially infer them using the wearable records.

q_{gender}	q_{height}	q_{bmi}	#
male	> 177.6	> 25	6
male	> 177.6	< 25	4
female	< 177.6	< 25	2
male	< 177.6	> 25	1
male	< 177.6	< 25	1
female	> 177.6	< 25	1

Table 3.2: Number of participants (#) in PMData who belong to a user group based on the answers to q_{gender}, q_{bmi} and q_{height}. Groups of size 1 contain uniquely identifiable individuals, who are prone to our attack. Parameters used for each prediction: steps, calories, and distance.

3.3 Membership inference on wearable data

The re-identification attacks introduced in the previous section are based on the assumption that the attacker is sure of her victim's presence in the targeted dataset. In this section, instead, we consider a scenario where the attacker does not know whether her target victim's data is present in the dataset. In this scenario, we devised a membership inference attack that aims to infer whether a particular individual's data is included in the dataset. This turned out to be not as effective as the other re-identification attacks, but still allows to determine if the target is present in the dataset with probability greater than random guessing.

This membership inference attack leverages additional records x' collected by a malicious actor, similarly to the re-identification attack in section 3.2.1.

Methodology The attack vector is as follows. First, the attacker the target's data x' to the closest participant in the dataset, adopting the same decision rule described in section 3.2.1. Then, she computes a score that estimates her confidence that the selected participant is actually her target victim. The confidence is inversely proportional to a score computed as

$$s(x_i, x') = \frac{1}{T'} \sum_{t'=1}^{T'} \min_t d(x_i[t], x'[t']). \tag{3.6}$$

where x_i is the closest time series to x' in the dataset. This score is compared to a threshold θ_s, which must be properly tuned.

In this attack, we do not evaluate only the probability of the attacker correctly inferring the presence of the target victim, but also whether the selected participant is actually the victim. In other words, the attack is considered successful only in the following cases:

- if the victim is correctly predicted as "present" *and* the closest time series in the dataset actually belongs to him;
- if the victim is correctly predicted as "not present" in the dataset.

Therefore, if the target victim is included in the dataset, the success rate of this membership inference attack is bounded by the success rate of the re-identification attack described in section 3.2.1.

Experimental evaluation Depending on whether the target participant is actually included or not, the detection rule can lead to different outcomes:

- *true positive* (TP), if the target is included and the rule correctly identifies him;
- *false negative* (FN), if the target is included and the rule fails to detect him;
- *true negative* (TN), if the target is not included and the rule concludes that he is actually not present;
- *false positive* (FP), if the target is not included and the rule detects one user as the target.

It is worth remarking that these definitions are different from the usual notions adopted in binary classification problems. Distinguishing between these outcomes allows us to estimate the success rate for the attack in the cases where the user is present or not in the dataset. Denoting with p_{in} the probability that the user is in the dataset, the success rates for the cases $p_{\text{in}} = 1$ and $p_{\text{in}} = 0$ are computed as

$$\Pr[\text{Success}|p_{\text{in}} = 1] = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.7}$$

and

$$\Pr[\text{Success}|p_{\text{in}} = 0] = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{3.8}$$

respectively. Using the law of total probability, any other case where $0 < p_{\text{in}} < 1$ can be expressed as a combination of the previous two as follows

$$\Pr[\text{Success}|p_{\text{in}}] = p_{\text{in}} \Pr[\text{Success}|p_{\text{in}} = 1] + (1 - p_{\text{in}}) \Pr[\text{Success}|p_{\text{in}} = 0]. \tag{3.9}$$

Figure 3.5 shows the role played by the threshold θ_s, which essentially represents the confidence required by the adversary to determine whether the target is present in the dataset, fixing the number of participants at $n = 15$. A low threshold implies that the attacker will conclude more often that the target is not in the dataset, leading to maximal accuracy for the adversary when the target is actually not present, i.e., $p_{\text{in}} = 0$.

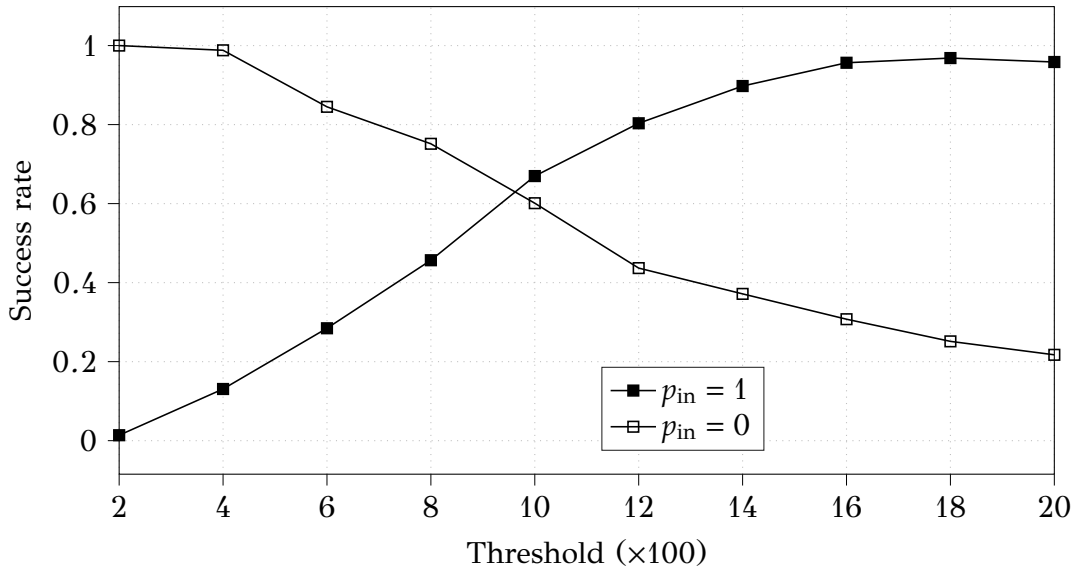


Figure 3.5: Success rate of the membership inference attack for varying threshold when the probability of the victim being included in the dataset is $p_{\text{in}} = 0$ and $p_{\text{in}} = 1$. The number of participants is kept fixed at $n = 15$. The threshold values were upscaled by a factor of 100 to make the axis more readable. Parameters used to link two records: steps and calories.

Conversely, a large threshold leads to a higher success rate when $p_{\text{in}} = 1$. In particular, when the threshold is sufficiently high, this attack is equivalent to re-identification via linking for $p_{\text{in}} = 1$. A balance between the cases $p_{\text{in}} = 0$ and $p_{\text{in}} = 1$ is given for a threshold around 10 (actually 0.1, since the threshold values are upscaled by a factor of 100).

After finding this suitable threshold value, we consider a more interesting case where the probability of the target being included in the dataset is $p_{\text{in}} = 0.5$. We test our method for a varying number of users in the dataset, comparing the performance of our attack with two naive rules, as shown in figure 3.6. The first one is *naive rejection*, which consists in always concluding that the user is not present in the database and has success probability $1 - p_{\text{in}}$. The second one is *naive guessing*, and consists in choosing with equal probability one of the users or "not present in the database". The success probability of such rule is $1/(n + 1)$ regardless of the value of p_{in} (it can be easily verified by applying the law of total probability). Our method shows better performance of both the naive rules in most cases. Naive rejection appears to provide close performance, but on the other hand it never allows to find the target user when he is actually present

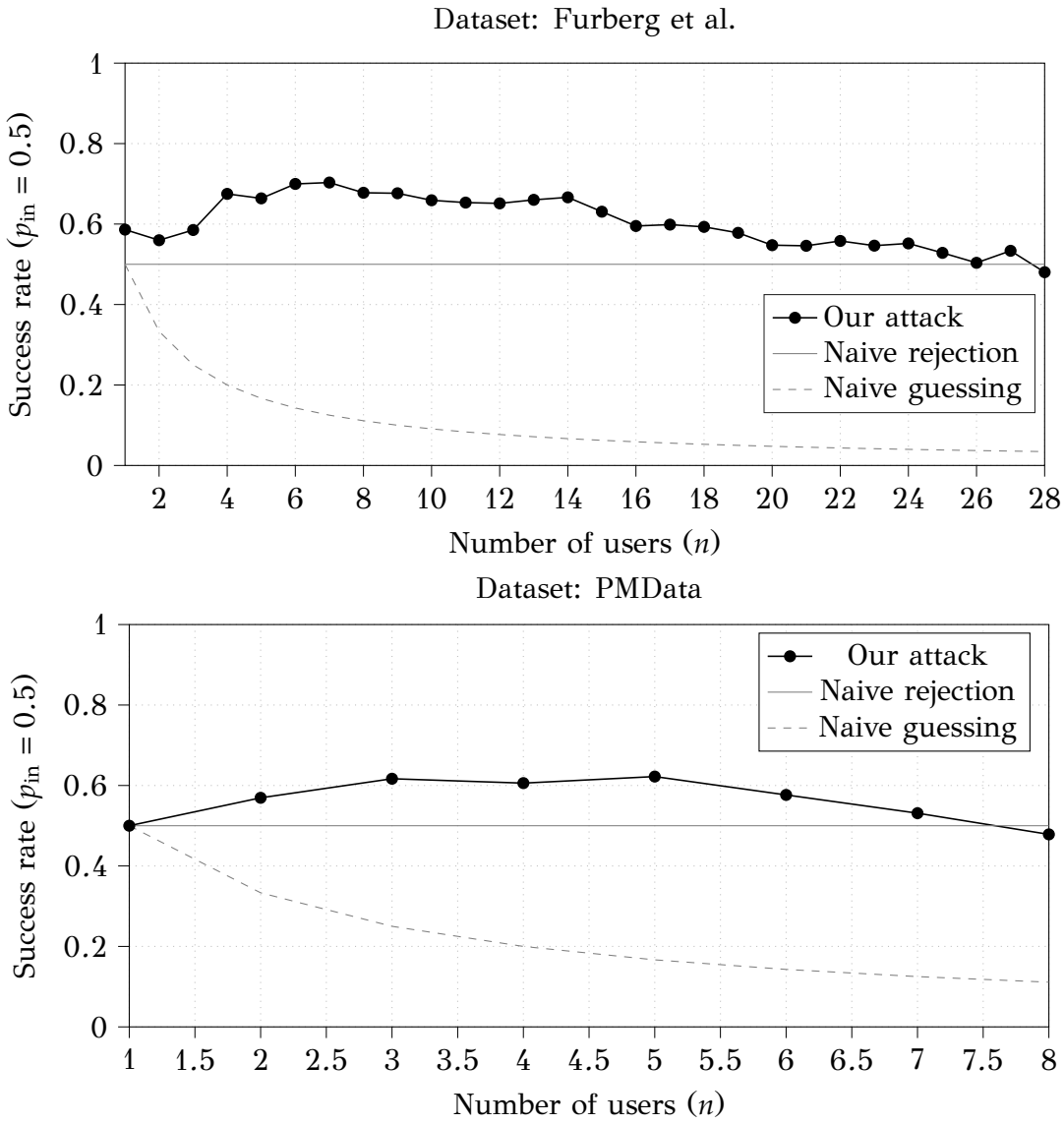


Figure 3.6: Success rate of our attack, estimated for varying number n of users and fixed threshold 10. The probability of the target being included in the dataset is fixed at $p_{in} = 0.5$. Parameters used to link two records: steps and calories.

3.4 Hourly records as a fingerprint

In all previous experiments, we assumed that the attacker attempts to re-identify or infer the presence of her target using daily activity records. This is a reasonable assumption, since the default option for most commercial wearables such as Fitbit is

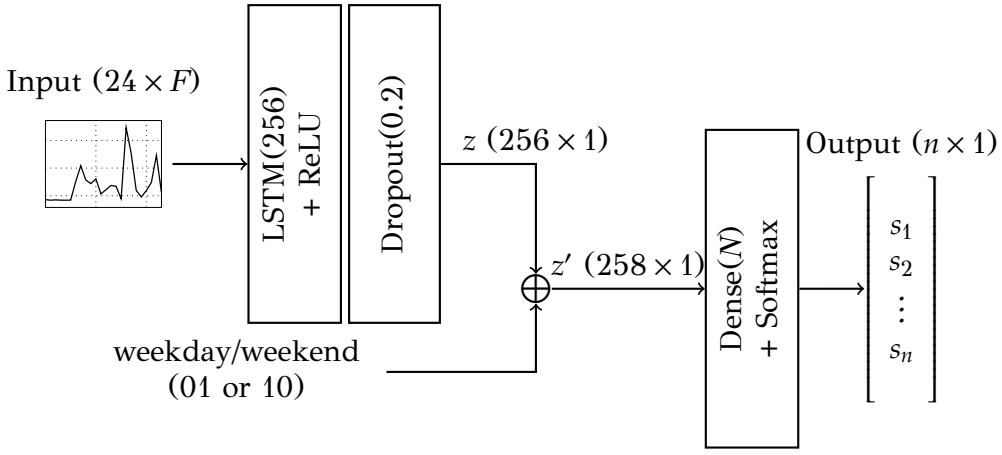


Figure 3.7: Architecture for re-identification based on 24-hour sequences of hourly-sampled wearable records. In our experiments, we used $F = 4$ features (steps, calories, distance, and average heart rate). An LSTM layer is employed to process the records as a sequence and produce a single 256×1 output. Two bits are concatenated to the output of the LSTM to model weekdays (01) or weekends (10), and processed to obtain the final prediction (a vector of n probabilities, $s_1, \ldots, s_n$, one for each participant).

to share daily aggregated information with friends or social media. However, another case that we consider is that the attacker has some knowledge about the routine of her target. In other words, do hourly-sampled records reveal insights on the target's routine?

In order to test this hypothesis, we consider time series data of hourly records, covering a day (24 hours) each, employing measurements of steps, calories, distance, and average heart rate. We used 80% of these samples to train a recurrent neural network model and tested the model on the remaining 20%. The model is based on an LSTM layers to process time series data. We incorporated information regarding whether the time series was recorded in a weekday or during the weekend. We used a standard one-hot encoding scheme for categorical information [103], where '01' represents a weekday and '10' represents a weekend. These additional values were concatenated to a latent representation of the time series following the application of the LSTM layer, as shown in figure 3.7. The output of the model is n scores, one for each user in a given dataset. The only dataset containing an adequate number of time series data for each user is PMData, so we use that to train and test our recurrent architecture.

Our model achieves a 93.5% de-anonymization accuracy when training and testing

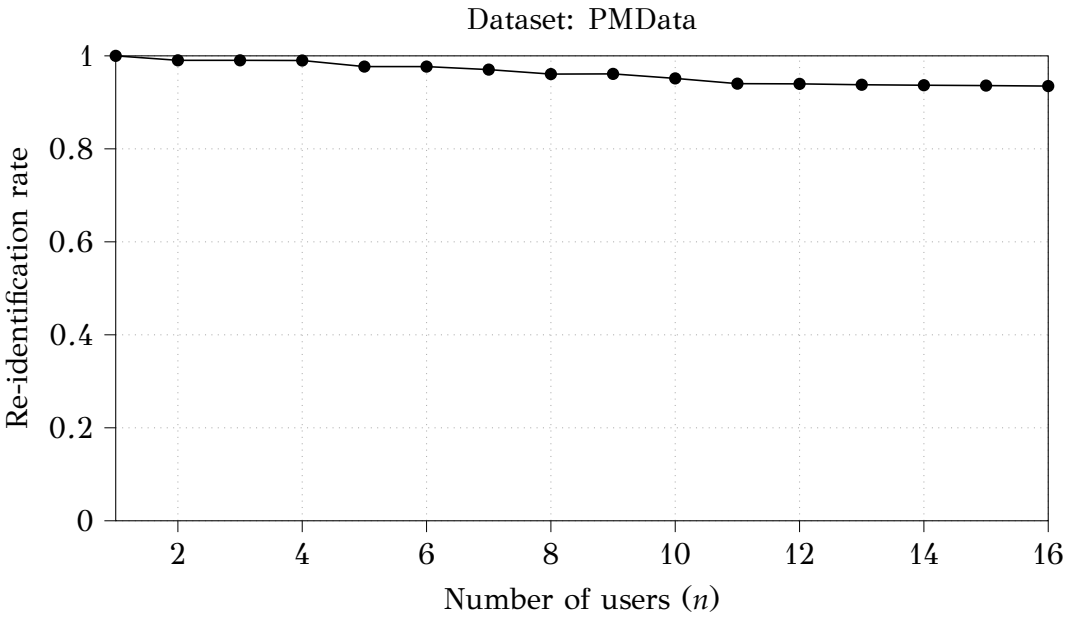


Figure 3.8: Re-identification rate based on 24-hour time series of hourly-sampled wearable records for a varying number of users n sampled from PMData. Features used for prediction: steps, calories, distance, average heart rate.

it on all the 16 members of PMData. Furthermore, we report the re-identification results for different numbers of participants, as shown in figure 3.8. We perform a Monte Carlo simulation, where for every number of users n, ranging from 2 to 15, we run 10 rounds of the simulation in which we select n participants at random every time. Then, we average the results for each value of n to get a final accuracy estimation. Our findings suggests that it may be possible to maintain high re-identification rate even for $n > 16$.

Additionally, we stress the fact that our results show the re-identification rate on single 24-hour time series records. If an attacker has access to more records, the probability of re-identification increases further.

3.5 Takeaways

- Enforcing k-anonymity on demographic and physical characteristics is not sufficient to protect the anonymity of participants in datasets of wearable data.

- An attacker can re-identify a target participant in a dataset of n participants by exploiting additional information that she has gathered. This could be either (i)

additional records that the target posted online or (ii) knowledge about personal characteristics of the target such as gender, height, and weight.

- Daily records of steps and calories collected by the same individual can be linked according to a minimum distance criterion. This is due to calories being computed as a function of steps and personal characteristics of the individual.
- Daily records of steps, calories, and distance can be used to determine personal characteristics of individuals in the form of binary information. Specifically, they can be used to determine gender, whether they are overweight or not, and whether their height is above or below average.
- Linking daily records can lead to a high probability of re-identification if the attacker is certain that her target is present in the dataset. In cases where the target's presence is not a certainty, the probability of an attacker inferring it is not as high.
- 24-hour time series of hourly-sampled activity records (steps, calories, distance, and average heart rate) constitute a fingerprint for a wearable device user.

Chapter 4
Publishing wearable data

In the previous chapter, we highlighted the potential risks associated with careless disclosure of wearable data. In this chapter, instead, we explore various solutions available in the academic literature that aim to mitigate these risks and protect the privacy of wearable data.

Our primary objective is to provide guidelines for safeguarding wearable data privacy that can be used to publish open datasets and make them publicly available for academic research. To this end, we examine several privacy protection techniques, and explore their applicability to wearable data. In addition, we apply these privacy protection guidelines to LifeSnaps, a real-world dataset consisting of data collected from 71 participants. We demonstrate that by applying these privacy protection measures, we can enhance the privacy protection of LifeSnaps participants compared to other public datasets.

The outcome of this chapter is a set of privacy protection guidelines that researchers can use to ensure the privacy and security of wearable data in academic studies. Our hope is that these guidelines will facilitate the creation of more open datasets that can be used to advance research in the field of health studies without compromising the privacy of individuals.

4.1 Privacy protection techniques

Before introducing our proposed guidelines, we survey different approaches proposed in literature to protect wearable data. We mainly focus on techniques that enforce privacy by removing or altering information, such as k-anonymity [121], differential privacy [27], privacy-preserving neural networks [124], and synthetic data. Other approaches rely on obfuscation techniques that protect data via cryptography such as homomorphic encryption [1], secure multi-party computation [49], and trusted execution environments [21]. However, these techniques do not completely eliminate the risk of a data breach or privacy violation [37, 39]. While they can protect data

from unauthorized access and manipulation, they do not prevent entities who have access to the data from abusing their access rights to extract sensitive information or to perform unwanted computations. As a result, they are not suitable for protecting data that needs to be made public.

***k*-anonymity** Many works proposed to use k-anonymity to protect wearable data [4, 66, 72, 76, 139]. In most of these researches, the threat model for re-identification assumes that the attacker can recognize her target only by demographic and physical attributes (e.g., gender, age, height, and weight). In other words, time series of activity records are only treated as sensitive information to be protected, but the possibility that they could also identify individuals is not considered. Indeed enforcing k-anonymity on demographic and physical attributes is certainly reasonable. However, this is not sufficient to guarantee the anonymity of a dataset participants, as we showed in chapter 3.

Differential privacy Differential privacy has also been largely studied in the context of wearable data. Most of the differentially private systems for wearables proposed in literature are based on local differential privacy (LDP), where data are protected and sanitized by the device owner before collection. Saifuzzaman et al. wrote an extensive literature review of existing studies [112]. However, in this chapter we focus mainly on approaches that researchers can use to protect data after collection. Decentralized anonymization at the user's side is covered in chapter 5.

Privacy-preserving neural networks A recently proposed approach to data sanitization utilizes machine learning, specifically neural networks. The core idea behind machine learning-based solution to privacy is that the train models should learn common patterns present in the data and preserve them, maintaining most of the utility. Many architectures designed in literature rely on the concept of adversarial machine learning, where two neural networks compete to achieve opposite goals and improve together. Specifically, one network aims to sanitize the data, while the other tries to infer sensitive information from the data. One pioneer work leveraging this concept was published by Tripathy et al. [124], which named this approach "privacy preserving adversarial networks". Malekzadeh et al. [77] applied similar approaches to time series of raw sensor data. The idea gained popularity, leading to follow-up research [13,78,106]. However, a main drawback of using neural networks for data anonymization is that these approaches do not provide any kind of privacy guarantees other than empirical results. Furthermore, being neural networks mainly black-box models, it is not clear how privacy is enforced and what information is removed. This approach, like any

other form of data protection, reduces the utility of the data. If no privacy guarantees are obtained in exchange, the utility loss is hard to justify.

Synthetic data In order to protect the privacy of individuals and prevent the risks of privacy breaches, many applications and research projects are turning to synthetic data. Synthetic data are essentially "fake" data that are generated artificially. While it is easy to produce synthetic data by randomly sampling from a probability distribution, the challenge lies in creating synthetic data that accurately reflect real-world samples.

This is especially important in health research where useful synthetic data should mirror the characteristics of actual people [18]. Considering machine learning applications, synthetic data could be used to train models that perform specific tasks, such as predicting calorie consumption based on steps and demographic information. In this case, the synthetic data should realistically represent the relationship between steps and calories.

To achieve this, several techniques have been developed to generate synthetic data based on real-world data. An approach that gained significant momentum is to use neural networks, such as generative adversarial networks (GANs) [41, 55, 83, 130] and variational autoencoders [9, 69]. These models can mimic the statistical properties of real-world data and generate synthetic data that are similar to the original data.

In the specific context of wearable data, Imtiaz et al. [52] used GANs with differential privacy to create synthetic Fitbit records. However, the usefulness of synthetic wearable data has not yet been fully evaluated.

4.2 Guidelines for publishing wearable data

After carefully considering the solutions proposed in the academic literature and comparing them with our findings, we developed a set of guidelines that researchers can follow when publishing datasets of wearable data [80]. The purpose of these guidelines is to mitigate the effectiveness of re-identification attacks while preserving as much useful information as possible. We also would like to stress the fact that our guidelines do not provide theoretical privacy guarantees, but can help mitigating various types of attacks that can be carried out against the dataset participants.

- *k-anonymity*. Albeit not sufficient by itself, applying k-anonymity is a necessary steps to mitigate privacy threats. k-anonymity should be enforced on the attributes that surely constitute quasi-identifiers. These include demographic information (such as gender, age, and country of origin) and physical characteristics (such as height). It is debatable whether weight can be considered a quasi-identifier, since contrarily to height it can significantly vary between different

periods in time. Nonetheless, being underweight or overweight could be a distinct characteristic. Therefore, it is a good idea to group people based on their body mass index (BMI).

- *Aggregation* Rather than publishing the characteristics of each individual, a sensible idea is to release them in an aggregated for, e.g., using tables or histograms. This mitigates re-identification attacks that match individuals based on such characteristics. However, the data curator should still carefully select the amount of aggregated information to release, since disclosing too many aggregated statistics may lead to the reconstruction of the original data [35].

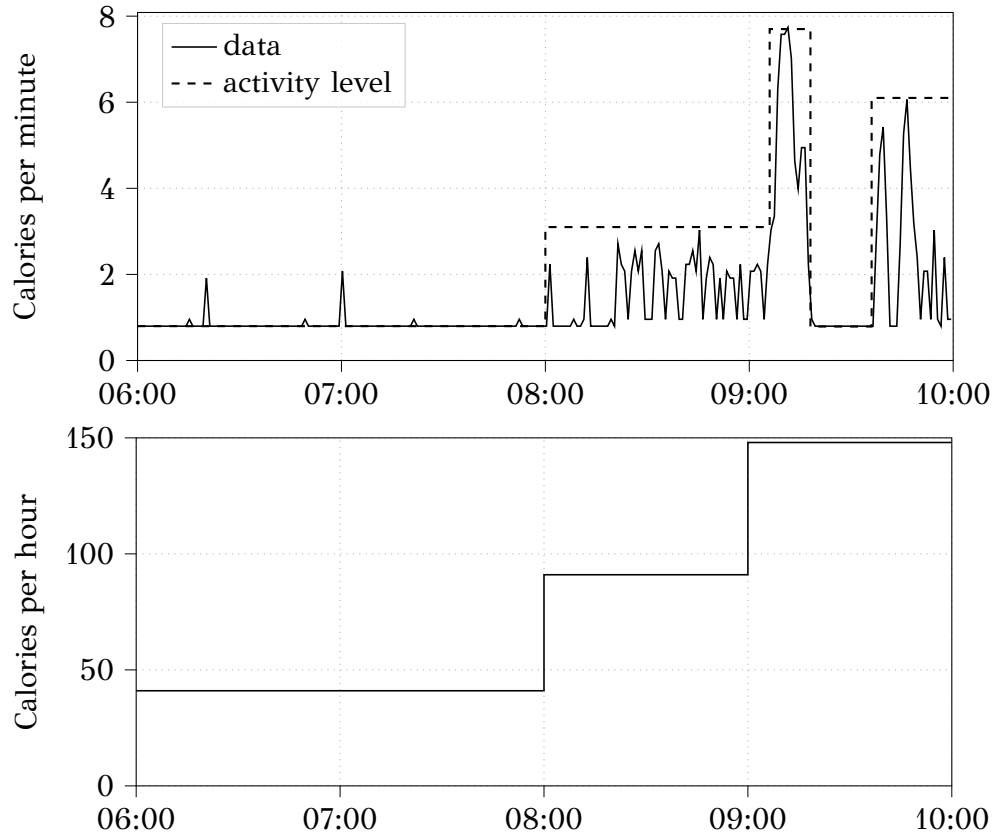


Figure 4.1: Morning routine of a user captured by calorie records, collected minute-by-minute and hourly. It can be determined in both cases that the user woke up around 08:00 AM. However, minute-by-minute records give more insights on the time spent active or at rest.

- *Resampling/subsampling*. Some time series data are recorded with high sampling rate, usually using the default settings of the device. For example, in Fitbit devices

Dataset	Quasi-identifiers	Re-identifiable
Furberg et al.	weight, BMI	13/35
PMData	age, gender, height, weight	16/16
Fitbit Connections	first name, height, weight	39/39

Table 4.1: Publicly available datasets with relative quasi-identifiers. The last column of the table shows the number of users who are characterized by a unique tuple of quasi-identifiers.These users can be re-identified by an attacker who can link these quasi-identifiers to their identity.

energy consumption and heart rate are sampled frequently, possibly revealing unique patterns as demonstrated in chapter 3. Aggregating values by hour or by day can help hiding such patterns. An example of subsampling where calorie records are aggregated by hour is shows in figure 4.1.

- *Quantization*. Collecting data with excessively high granularity may facilitate attacks in which an adversary knows the exact number of steps taken by her target on a specific day. This may be avoided through quantization, by binning the data into predefined values (e.g., 12879 → 12500).

In general, wearable data records should be published in accordance with the *principle of data minimization* [11]. This principle can be summarized as follows: "The amount of disclosed personal data should be limited to what is necessary in relation to the purpose for which they are collected." In other words, if a dataset is created with a specific objective in mind, it should only contain the minimum amount of information required to achieve that objective. This means that unnecessary personal information should be removed to protect the privacy of the individuals involved. Conversely, if the data are collected "in the wild" – i.e., without a specific purpose – publishers should still privilege the anonymity of the participants over the dataset utility. Finally, the most effective way of protecting individuals in a dataset is gathering many participants, making it harder for an adversary to find a specific target.

4.3 Privacy analysis of public datasets

Due to the lack of available guidelines, most of the existing public datasets are not properly protected against privacy threats. We observed that the majority of them does not apply basic anonymization of demographic and physical quasi-identifiers. This makes them prone to re-identification attacks that rely on this information.

Furberg et al.'s dataset contains weight and BMI updates for 13 out of 35 users. These records are more specific then just a single reported weight measurement and it

is more likely for an adversary to connect this information to an individual's identity. Furthermore, pairs of weight and BMI can be used to calculate the height of the participants.

PMData includes a spreadsheet with personal information for each participant, including age, gender, height, and weight, without any form of generalization. Although weight changes over time and thus may not be useful to identify participants, all athletes (16 out of 16) can be uniquely characterized by their combination of age, gender, and height.

Fitbit Connections even contains the names of the participants (although few of them utilized pseudonyms). Furthermore, information about their height and weight is made available. Even if the name is not considered for participants who used a pseudonym, all the individuals in the dataset (39 out of 39) are uniquely characterized by a tuple of quasi-identifiers.

In summary, all the public datasets of wearable data are prone to the linking attack presented in the *k*-anonymity section of this thesis (see table 2.2).

4.4 Anonymization of LifeSnaps

Our guidelines have been applied to LifeSnaps [132], a multimodal dataset of Fitbit records, which was collected and published by the RAIS consortium. The dataset includes records from 71 participants recruited from four countries: Cyprus, Greece, Italy, and Sweden. For each participant, the dataset contains information on their gender and body mass index (BMI). Additionally, aggregated information on age group, height, and education level is also available. The dataset stores approximately two months (or 60 days) of activity records, aggregated hourly, for each user.

Before collecting the data, we made a commitment to protect the privacy and sensitive information of the participants. Therefore, we took great care to thoroughly anonymize the dataset before publication. In doing so, we adhered to the following principles: (i) minimizing the likelihood of successful re-identification of users by real-world adversaries, (ii) retaining as much data as possible that would be useful to researchers and practitioners, (iii) following the principles and recommendations of GDPR regarding the handling of personal information, and (iv) adhering to established anonymization practices and principles.

We enforced 2-anonymity on the dataset with respect to demographic and physical characteristics, meaning that for each participant in the dataset there was at least another with the same characteristics. To achieve 2-anonymity, we removed all individual-level details from the dataset, except for gender and BMI. The personal information that was removed from the participants' details was presented in the form of histograms, as shown in figure 4.2. Additionally, we removed all traces of infor-

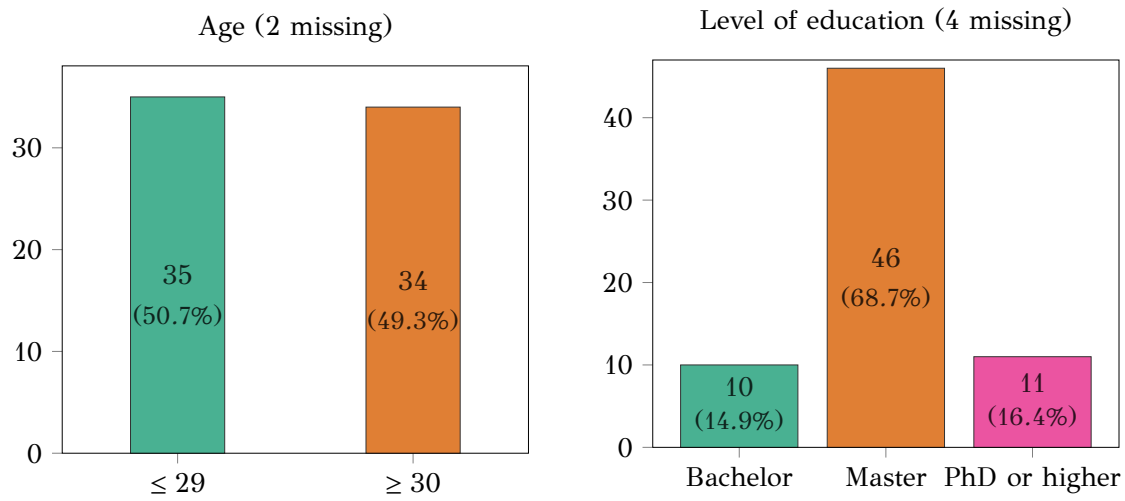


Figure 4.2: In LifeSnaps, details about age and education were removed for individual participants to achieve 2-anonymity, but were still reported in aggregated form as histograms. This allows to disclose the general demographics of the studied population without revealing personal information.

mation that could identify the participants such as records of specific activities (e.g., "canoeing") or information that was not reported in English.

4.5 Takeaways

- Due to the lack of established guidelines, existing public datasets of wearable records contain plenty of potentially identifying information.
- Many approaches have been proposed in literature, based on k-anonymity, differential privacy, sanitization via privacy-preserving neural networks, and synthetic data generation.
- Our own set of guidelines complement the approaches proposed in literature and are tailored specifically for wearable data. These include the use of k-anonymity on demographic and physical information, aggregation of sensitive attributes, subsampling of time series data, and quantization for records with high granularity.
- Our guidelines have been applied to LifeSnaps, a dataset of wearable data collected by 71 participants from 4 different countries.

Chapter 5
Decentralized solutions for wearable data collection

In the previous chapter, we proposed guidelines for protecting wearable data privacy before publication. However, these guidelines require participants to entrust their private data to a curator, and may still be susceptible to re-identification attacks.

In this chapter, we propose decentralized solutions for protecting wearable data during the collection phase, which are enforced directly on the appliances of the participants, thus removing the need for a trusted curator. Many works use the term "decentralized privacy solutions" referring to decentralized storage platforms [26, 75, 125], often built on blockchain technology [118], with restricted access control. These platforms only safeguard privacy by blocking unauthorized access to the data. However, this prevents the extraction of any valuable insights from them. Our solutions, on the other hand, are based on the concept of *differential privacy*, which provides theoretical privacy guarantees while allowing useful information to be extracted from the data. To design specific solutions that preserve the utility of the collected wearable data, we explore two use cases. In the first case, we consider wearable data collected in the context of comparative studies between groups of participants. In the second case, we examine wearable data collected for training machine learning models.

Our decentralized approach to protecting wearable data privacy potentially facilitates the collection of wearable data in an online setting, where establishing a relationship of mutual trust with participants can be difficult. This could make data collection more cost-effective and efficient, and enable more widespread data sharing for academic research. Overall, our proposed decentralized solutions offer a practical and theoretically sound way to protect the privacy of wearable data during the collection phase, paving the way for more effective and secure data sharing in the future.

5.1 Crowdsourcing setting

Online crowdsourcing potentially constitutes a cost-effective solution to facilitate health studies based on these devices, which are typically expensive in terms of both time and resources. By using online crowdsourcing platforms to recruit participants and gather data, researchers can select candidates who already own a suitable wearable tracker, saving them the need to buy new devices and meet participants in person.

In the rest of this chapter, we assume that our decentralized approaches are employed in online crowdsourcing platforms. Users of commercially available fitness trackers connect to these platforms and submit their data to an analyst, either voluntarily or under compensation. Online crowdsourcing is already used to collect wearable data from users who already own a fitness tracker. Both the Furberg et al. and Fitbit Connections datasets were collected via online crowdsourcing platforms (Amazon Mechanical Turk and OpenHumans, respectively).

However, in our case we propose our own design for a crowdsourcing platform [82], since existing ones do not meet the privacy requirements of our decentralized solutions for data collection.

More specifically, the requirements for the online crowdsourcing platform are as follows:

1. *Anonymity*: The analyst – and any other entity who has access to the data – must not be able to link data points to a specific participant.

2. *Quality*: The analyst must be able to use the collected data to carry out a certain task within an acceptable margin of error.

3. *Accountability*: The analyst must be able to reward participants when they send their data. Conversely, a participant who does not submit any data should not be rewarded.

As outlined in the introduction, the tasks that an analyst should be able to undertake are making a comparative study between two groups of users and to train machine learning models based on the received data.

5.2 Comparative studies with differential privacy

In this section we demonstrate how our proposed platform setting can be used to crowdsource wearable data under local differential privacy (LDP) and enable comparative studies based on the collected data. As mentioned in chapter 2, comparative studies are the most prominent application of wearable data in health research. Using a crowdsourcing platform increases the amount of participants that researchers can

reach and allow to sample from a diverse population, making the results of the studies more reliable.

Crowdsourcing platform design Our proposed solution involves multiple participants and an analyst communicating using a third-party server as intermediary, as depicted in figure 5.1. The communication pipeline between these actors can be summarized as follows. The analyst recruits n users as participants in a health study. Both the participants and the analyst connect to a crowdsourcing platform, which is represented as a server. Upon sign up, the server assigns a unique user identifier (UID) to each participant. At the beginning of the experiment, the analyst generates an asymmetric key pair. She keeps the secret key sk for herself and distributes the same public key pk to each user. Individuals locally randomize their reports and encrypt them with pk. Then, they submit the encrypted data to the server along with their UID. Afterwards, the server replaces UIDs with random report identifiers (RIDs) and sends the (RID, encrypted data) pair to the analyst. The server must generate a new RID for each new submission. After decrypting a record and verifying its integrity, the analyst sends a (RID, reward) pair to the server. The server, in turn, forwards the reward to the user with UID matching the RID.

How privacy is achieved In our solution, both the analyst and the third-party server are not able to compromise the anonymity of the participants under the honest-but-curious model, i.e., assuming that they do not actively conspire against the users.

- The third-party server knows which user (identified by the UID) has submitted a given report. However, since the report is encrypted, the server is not able to see its content.

- The analyst is able to observe the content of a report, since she owns the private key sk. She does not know which user has submitted such report, since it was forwarded by the third-party server and associated with an RID. Furthermore, she is not aware of whether two distinct reports belong to the same user.

- Reports are randomized with ε-LDP to prevent the analyst from recognizing the user based on some "fingerprint" contained in the data. As long as a suitable value of ε is chosen, participants cannot be re-identified. This is due to LDP providing statistical indistinguishability between randomized records and is explained more in details in section 5.2.4.

- All participants should use the same public key pk to encrypt their traffic, so that this does not become an identifier. If the study involves comparing two groups

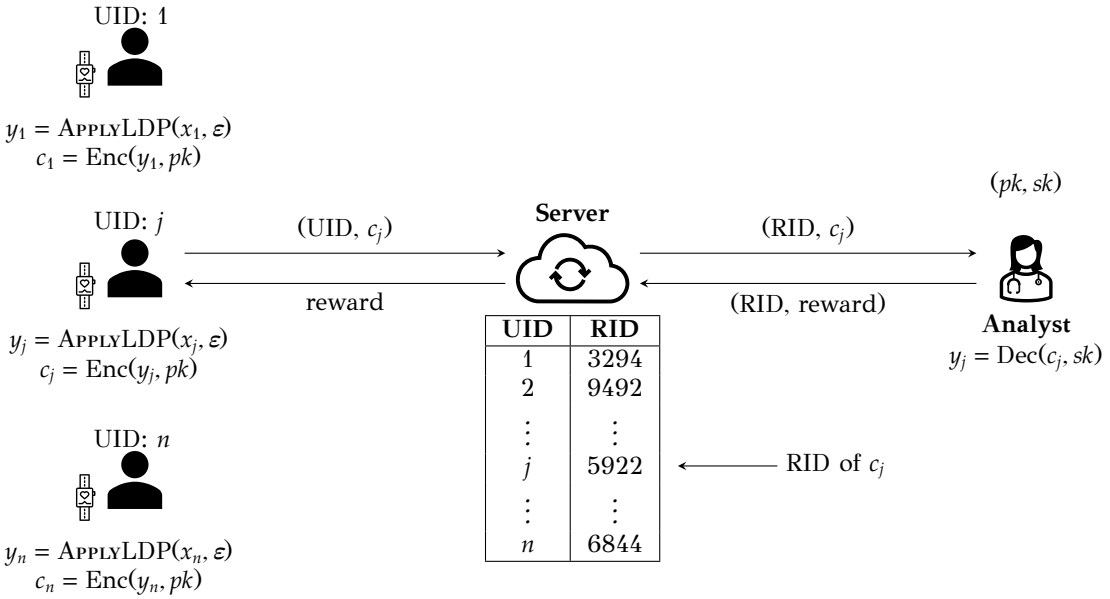


Figure 5.1: Design of a crowdsourcing platform that guarantees anonymous reporting under local differential privacy (LDP). Users submit their wearable IoT data once a day. A participant with user identifier (UID) j randomizes his daily report using ε-LDP and encrypts it with a public key pk. The participant sends the pair (UID, c_j) to a third-party crowdsourcing server, which assigns a random report identifier (RID) to c_j. The server forwards the pair (RID, c_j) to the analyst, who decrypts the report using a secret key sk and sends back a reward for the corresponding RID. The server then forwards the reward to the appropriate user. This pipeline guarantees user anonymity, unless the analyst and the server work together to compromise a user.

of participants, as it is typically done in randomized control trials, each group may use a different public key.

Independent reports and privacy budget Besides guaranteeing the sender anonymity for individual submissions, an important property of our three-party scheme is that reports submitted by the same user can be considered "independent". Not disclosing UIDs to the analyst also enables users to send multiple ε-DP reports without allocating more privacy budget, as long as the RID is changed. If the same user were to submit multiple reports under a same identifier, he should divide his budget between the privacy reports. This means that if he wanted to allocate an overall privacy budget of ε for L reports, he should apply LDP with budget ε/L to each report. However, if the RID changes, the analyst is not able to tell that two reports have been submitted

by the same user. Therefore, each report can be perturbed with budget ε. Indeed, the requirements are satisfied only if both the server and the analyst are either completely honest or "honest-but-curious", meaning that they follow the rules while trying to lawfully glean as much information as possible. If the third-party server reveals the actual UID to the analyst or does not change the RID over multiple submission, then the reports would not be independent anymore. Thus, the analyst would be able to glean more information on the users.

Collected dataset Assuming that the analyst and the server do not conspire, the final collected dataset would be a time series of sets ranging over T days

$$D = (D[1], \ldots, D[t], \ldots, D[T]). \tag{5.1}$$

Each set $D[t]$ is an unordered collection of n anonymous reports

$$D[t] = \{y_1[t], \ldots, y_n[t]\}. \tag{5.2}$$

In comparative studies and randomized control trials, two separate datasets D_A and D_B should be collected, one for the experimental group and one for the control group. The reports are forwarded to the analyst in random order, ensuring that the i-th report on different days t and t' does not pertain to the same participant. Collecting unordered reports in an unordered manner is crucial for protecting privacy. If an attacker were to obtain a complete time series of T reports from the same target, the privacy protections achieved through randomization would rapidly vanish.

We provide a simple example to illustrate this point. Suppose the dataset instead consists of ordered reports, meaning that $y_i[1], \ldots, y_i[T]$ belong to the same participant. If an attacker is aware that her target generally takes a similar number of steps per day, she may conclude that the sequence with the least variance between step reports belongs to the target.

While essential for safeguarding privacy, a dataset consisting of unordered records inevitably imposes certain limitations. Specifically, any analysis that relies on multiple records from the same user is not feasible. On the other hand, aggregated metrics that involve all users within the dataset are permissible. For instance, it is possible to monitor the activity trend of the entire group as a whole, but monitoring the activity of individual users is not permitted. These limitations align with the core principle of differential privacy mechanisms, which allow analysts to compute aggregated queries while preventing individuals from being singled out.

5.2.1 Methodology

We assess the preservation of data quality in LDP by evaluating the accuracy of an analyst's estimation of metrics of interest that are commonly used in comparative studies based on wearables. Specifically, we derive estimators for calculating the sample average, inverse cumulative distribution function (ICDF), and p-value of t-tests based on noisy samples $y_1, \ldots, y_N$. Daily calculation of these metrics enables monitoring of participant progress during rehabilitation or physical activity interventions. Comparing two populations using the sample average can help assessing the effectiveness of certain strategies, such as encouraging participants to take more steps. Statistical significance of such comparisons can be determined using the p-value. Finally, the ICDF $Q(x)$ estimates how many people have taken more than a given number of steps, showing if they met a certain fitness goal, e.g., $x = 10,000$.

Estimators under LDP The introduction of noise in reported records by LDP must be taken into account when estimating the metrics of interest. In the following paragraphs, we define and justify the use of estimators for the sample average, ICDF, and p-value. Our objective for the sample average and ICDF is to obtain an unbiased estimation, meaning that the estimation based on the noisy records should be equal, on average, to the metric computed on the original records. In contrast, for the p-value, we prefer an estimation that is biased towards a specific type of error, as we will explain later in this section. All the estimators are derived for single-valued records but can be applied separately to each measurement for records containing multiple values (such as steps, calories, distance).

- *Sample average*: The sample average of the original samples $x_1, \ldots, x_n$ is simply $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$. The r.v. Y_i, representing the i-th randomized report, has mean $\mathbb{E}[Y_i] = x_i$ for both the Laplace and Piecewise mechanisms. Therefore, it holds

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} Y_i] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i] = \frac{1}{n}\sum_{i=1}^{n} x_i = \mu \tag{5.3}$$

  meaning that

$$\hat{\theta}(\mu) = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5.4}$$

  is an unbiased estimator for the sample average. This holds both when x is a scalar or a vector of parameters. We denote the sample average for a feature f at day T as $\mu_f[t]$ and its estimator as $\hat{\theta}(\mu_f[t])$.

- *Inverse Cumulative Distribution*: The empirical ICDF of $x_1, \ldots, x_n$ is computed for

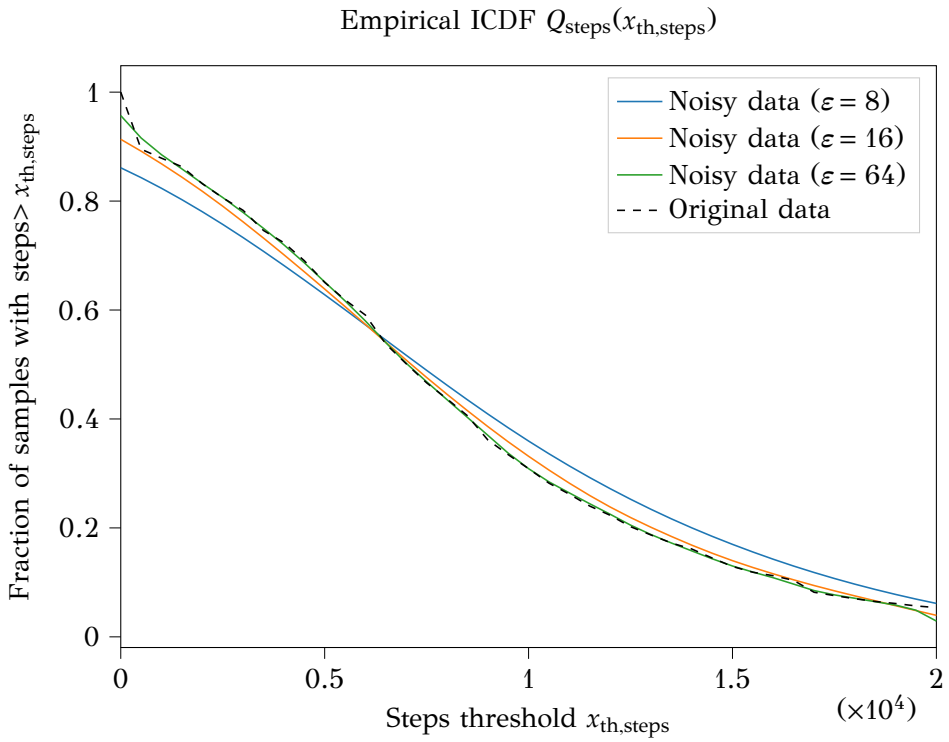


Figure 5.2: Example of empirical ICDF estimation for different values of ε. Evaluating the ICDF allows to count how many participants achieved a certain step goal.

each scalar value $x \in \mathbb{R}$ that a certain parameter can take as

$$Q(x) = \frac{1}{n} \sum_{i=1}^{n} \chi\{x_i > x\}, \tag{5.5}$$

representing the fraction of participants for whose reported value x_i is above a given threshold x. Under the observations $Y_1 = y_1, \ldots, Y_N = y_n$, the empirical ICDF can be estimated as

$$\mathbb{E}[Q(x)|Y_i = y_i] = \mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} \chi\{X_i > x\}|Y_i = y_i] \tag{5.6}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\chi\{X_i > x\}|Y_i = y_i] \tag{5.7}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \Pr[X_i > x|Y_i = y_i]. \tag{5.8}$$

For the Laplace mechanism, because of the additive relation $Y_i = X_i + Z_i, Z_i \sim \text{Lap}(0, \Delta/\varepsilon)$, we have that $X_i = Y_i - Z_i$. Thus, under the observations $Y_i = y_i,\ i = 1, \ldots, n$, the estimated empirical ICDF becomes

$$\hat{\theta}(Q(x)) = \frac{1}{n}\sum_{i=1}^{n} \Pr[y_i - Z_i > x] \tag{5.9}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \Pr[Z_i < y_i - x] \tag{5.10}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \begin{cases} \frac{1}{2}e^{\varepsilon\frac{y_i - x}{\Delta}} & \text{if } y_i \leq x, \\ 1 - \frac{1}{2}e^{\varepsilon\frac{x - y_i}{\Delta}} & \text{otherwise.} \end{cases} \tag{5.11}$$

Figure 5.2 shows that eq. 5.11 can effectively be used to estimate the empirical ICDF. We use $Q_f(x_{\text{th},f})[t]$ to indicate the empirical ICDF at day t for feature f evaluated w.r.t. the threshold $x_{\text{th},f}$. To estimate the actual number of participants (*count estimate*) who are above the threshold $x_{\text{th},f}$ at day t, we can simply multiply the value of the ICDF by n, i.e.,

$$\hat{\theta}(n \cdot Q_f(x_{\text{th},f})) = n \cdot \hat{\theta}(Q_f(x_{\text{th},f})). \tag{5.12}$$

- *Independent t-test*: Running a t-test requires to calculate the sample mean and variance for two groups of participants. To estimate the p-value from the anonymous reports, we simply run a normal t-test on the noisy samples. We first estimate the t_{stat} statistic based on two collections of anonymous reports $y_A^{(1)}, \ldots, y_A^{(n_A)}$ and $y_B^{(1)}, \ldots, y_B^{(n_B)}$, with $n_A + n_B = n$, as

$$\hat{\theta}(t_{\text{stat}}) = \frac{\hat{\theta}(\mu_A) - \hat{\theta}(\mu_B)}{\hat{s}/\sqrt{n}}, \tag{5.13}$$

where $\hat{s}$ is the overall sample standard deviation. Based on $\hat{\theta}(t_{\text{stat}})$, we compute the corresponding p-value $\hat{\theta}(p)$. Applying LDP does not introduce a bias in the mean values μ_A and μ_B, since they can increase or decrease with equal probability. This implies that the difference $\mu_A - \mu_B$ is also estimated without bias. On the other hand, the sample variance is increased by the variance of the noise, which may lead to an overestimation of the p-value. However, it is not worth compensating for the additional variance, as overestimating the p-value is preferable to an underestimation.

Quality metrics To assess the accuracy of estimated metrics of interest, we compare the values obtained by calculating them on the original and randomized data. For numerical values such as sample average and ICDF, we utilize the RMSE to make such comparison. For the p-value, instead, we are only interested on whether the obtained results are above or below the significance threshold a. Ideally, we would like the original and randomized data to yield the same results in term of significance. To measure how frequently this happens, we compute the *agreement rate* between t-tests.

For the sample average μ_f and the ICDF $Q(x_f)$, we would like to estimate the standard error on such metrics. If the estimators are unbiased, the standard error can be estimated by computing the *root mean square error* (RMSE) between the estimated and true metrics. The RMSE for the sample average is computed over T days for a specific metric f (e.g., steps). Let $\hat{\theta}(\mu_f[t])) \in \mathbb{R}$ be the estimated sample average at day t, and let $\mu_f[t]) \in \mathbb{R}$ be its true value, calculated without applying LDP noise. The RMSE over T days is calculated as

$$\text{RMSE}(\mu_f, T) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\mu_f[t]) - \hat{\theta}(\mu_f[t]))\right)^2}. \tag{5.14}$$

We choose RMSE over mean absolute error (MAE), used in other works [112], since RMSE penalizes sporadic large errors, and in randomized controlled trials consistently low errors are desirable. When comparing different features, we normalize the RMSE to express the error in a percent form. This metric is called normalized RMSE (NRMSE) and is computed as

$$\text{NRMSE}(\mu_f, T) = \frac{\text{RMSE}(\mu_f, T)}{x_{\max,f} - x_{\min,f}} \tag{5.15}$$

for the sample average. The values $x_{\min,f}$ and $x_{\max,f}$ are established before the experiment. These are also used to determine the sensitivity of the LDP mechanisms. The RMSE for the ICDF of a feature f over T days is evaluated in a similar fashion

$$\text{RMSE}(Q_f(x_{\text{th},f}), T) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(Q_f(x_{\text{th},f})[t] - \hat{\theta}(Q_f(x_{\text{th},f})[t]))\right)^2}. \tag{5.16}$$

However, in our experiments, we are more interested in measuring the error on the actual count estimate $n \times \text{ICDF}$, computed as

$$\text{RMSE}(n \cdot Q_f(x_{\text{th},f}), T) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(n \cdot Q_f(x_{\text{th},f})[t] - \hat{\theta}(n \cdot Q_f(x_{\text{th},f})[t]))\right)^2}. \tag{5.17}$$

	$p < a$	$p \geq a$
$\hat{\theta}(p) < a$	Agreement (both tests show significant difference)	Type I error
$\hat{\theta}(p) \geq a$	Type II error	Agreement (both tests show no significant difference)

Table 5.1: Different types of agreement and errors in t-tests under LDP. A standard threshold value is $a = 0.05$, which implies 95% confidence.

We consider the RMSE on the ICDF as the NRMSE of the count estimate, since it is divided by n, i.e.,

$$\text{NRMSE}(n \cdot Q_f(x_{\text{th},f}), T) = \text{RMSE}(Q_f(x_{\text{th},f}), T). \tag{5.18}$$

The agreement rate is an accuracy metric that we use to determine the reliability of t-tests under LDP. In principle, if p and $\hat{\theta}(p)$ are the p-values computed on the original and noisy samples, respectively, we would like them to be both above or below the significance threshold a, i.e., $\hat{\theta}(p) < a \Leftrightarrow p < a$. The *agreement rate* over n_{trials} trials with p-value threshold a is as

$$\frac{1}{n_{\text{trials}}} \sum_{\nu=1}^{n_{\text{trials}}} \chi\{p^{(\nu)} < a \wedge \hat{\theta}(p^{(\nu)}) < a\} + \chi\{p^{(\nu)} \geq a \wedge \hat{\theta}(p^{(\nu)}) \geq a\}, \tag{5.19}$$

i.e., the percentage of test pairs that yield the same result. This represents an indicative value for the probability of two tests having the same significance. When the two t-tests are not in agreement, we distinguish between two types of error, as summarized in table 5.1: the type I error (false positive) occurs when $p < a$ but $\hat{\theta}(p) > a$, while type II error occurs in the opposite scenario. While t-tests can demonstrate the difference between 2 populations (if $p < a$), they cannot disprove such difference (if $p > a$). In other words, it cannot be concluded that 2 groups are statistically similar by running a t-test. Therefore, type I error is less desirable, since it means that we accidentally conclude that the two populations are significantly different, while in reality this is not the case. For this reason, having an estimator that overestimates the p-value is preferable. A systematic overestimation does not necessarily reduce the agreement rate, but rather makes type II errors more frequent and type errors I less frequent. This is also confirmed by our results, which show that our p-value estimator consistently achieves high-rate agreement when $p > a$ on the original data.

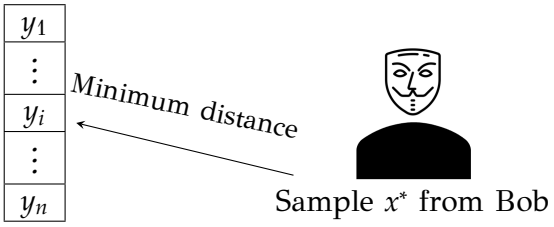


Figure 5.3: Linking attack considered in our evaluation. The adversary (Eve) aims to re-identify her target (Bob) by leveraging the original sample x^* and comparing it to the anonymized records $y_1, \ldots, y_n$.

Re-identification attack The purpose of applying LDP is to protect participants against re-identification, hence guaranteeing their anonymity. To determine the level of protection assessed by an LDP mechanism, we study its effectiveness against the following threat model, depicted in fig. 5.3: we assume that the adversary knows the original record x^* produced by her target on a certain day, and that she has access to the anonymous reports $y_1, \ldots, y_n$ from all the participants of a certain group. Therefore, a naive guessing approach would yield a $1/n$ re-identification probability. Contrarily to membership inference attacks, the adversary is assumed to be sure of the target's presence in the dataset of anonymous records. We also assume that the details of the LDP mechanism are known to the adversary. In other words, the attacker knows whether the Laplace or Piecewise mechanism was used to anonymize the data, and with what privacy budget.

Arguably this is the strongest attack that can be performed against a dataset collected under LDP, as we illustrate in section 5.2.4.

Indeed, this threat model is unrealistic. If the adversary already knows the original records, finding the corresponding anonymous report will not provide her with any additional information. Practical linking attacks leverage prior knowledge of the adversary about the target (e.g., "I know that the target is very active") or approximate information about a specific day (e.g., "On that date, the target ran a marathon"). Another strategy may consist in comparing (steps, calories) pairs to find individuals with similar height and weight, since these characteristics are used to estimate calories from steps. However, the threat model studied in this paper is stronger than most practical linking attacks. Therefore, its success rate can be considered an upper bound to the actual attack vectors that an adversary may adopt.

Once the adversary has access to x^* and $y_1, \ldots, y_n$, she needs a criterion to determine which report was most likely obtained by randomizing x^*. Intuitively, due to how the Laplace and Piecewise mechanisms are design, the "closest" report to x^*

is also the most likely to be its noisy counterpart. When the reports consist of a single feature, the optimal choice for the adversary is simply choosing the report that minimizes $|y_i - x^*|$. The measure of closeness that we adopt is the Euclidean distance between the original and noisy record with scaled features. Formally, the most likely report $\hat{y}$ that an adversary can choose is

$$\hat{y} = \arg \min_{y_i, i=1,\dots,n} \sum_{f=1}^{F} \left(\frac{|y_{i,f} - x^*_f|}{x_{\max,f} - x_{\min,f}} \right), \tag{5.20}$$

where $f = 1, \dots, F$ is the feature index. Each feature is scaled w.r.t. the sensitivity $\Delta_f = x_{\max,f} - x_{\min,f}$ since the amount of noise is proportional to the sensitivity. This criterion is optimal for the Laplace mechanism according to maximum a posteriori probability (MAP). For the Piecewise mechanism, the optimal decision is

$$\hat{y} = \arg \max_{y_i, i=1,\dots,n} \sum_{f=1}^{F} \chi \left\{ y_{i,f} \in \left(L(x^*_f), R(x^*_f) \right) \right\}, \tag{5.21}$$

i.e., choosing the report with most features in the high-density regions $(L(x^*_f), R(x^*_f))$. However, in most practical cases, this is equivalent to the minimum distance criterion. Therefore, in our experiments we adopt the criterion described in eq. 5.20, since it is faster to evaluate, and thus more suitable for Monte Carlo experiments.

5.2.2 Experimental results

We test the effects of LDP on LifeSnaps. The experiments reported below require both an adequate number of participants and a large number of records per participant. To our knowledge, LifeSnaps is the only dataset that satisfies both requirements. We vary the number n of participants from 1 to 71, and the privacy budget ε from 1 to 64. For each (n, ε) pair, we run a Monte Carlo experiment of $n_{\text{trials}} = 100$ iterations, where each iteration is as follows:

- We select n participants uniformly at random from the dataset;
- We apply the chosen LDP mechanism (Laplace or Piecewise) with budget ε;
- We compute the metrics of interest.

Metrics of interest are averaged across the n trials to produced the final reported values. In order to provide a sensible visualization of our results, we show how the privacy budget impacts each metric in two cases: (i) fixed number of participants $N = 30$ and variable privacy budget ε, and (ii) variable number of participants and

	$x_{\min,f}$	$x_{\max,f}$
Steps	0	20000
Calories	0	6000
Distance (m)	0	15000

Table 5.2: Minimum and maximum values chosen for all features. Each features is clipped in the interval $[x_{\min,f}, x_{\max,f}]$ before applying LDP. This allows to compute the sensitivity for the LDP mechanisms.

fixed privacy budget $\varepsilon = 8$. When measuring the error on aggregate metrics and agreement on t-tests, we report only the outcomes obtained for the step count, since that is the most widely-used features in fitness studies [7, 70]. While estimating the success rate of linking attacks, we consider the combination of steps and calories, since we have shown that these attributes are highly identifying. Other features yield similar results.

Both the Laplace and Piecewise mechanisms require the input to be bounded in a range $[x_{\min}, x_{\max}]$ to calibrate the noise. The amount of randomness to be introduced depends also on the width of this range, therefore, this cannot be too large. Thus, we clip input features in bounded intervals according to Table 5.2.

Sample average The RMSE is calculated on aggregated metrics by taking into account both original and anonymized samples across the 64 days in the dataset. Figure 5.4 illustrates the RMSE between the true sample mean μ_{steps} for steps and the estimate $\hat{\theta}(\mu_{\text{steps}})$, for various privacy budgets and numbers of participants. Increasing ε and/or n reduces the RMSE, which is expected since a higher number of records reduces the variance for the sample average estimator.

The Piecewise mechanism introduces less error than the Laplace mechanism at the same level of privacy budget. When the number of participants is $n = 30$ or higher, the Laplace mechanism introduces less than 600-steps error for $\varepsilon \geq 8$. This error level is acceptable, accounting for approximately 3% of the overall range $[0, 20000]$. On the other hand, the Piecewise mechanism reaches the same utility at $\varepsilon = 4$. This factor should be considered when selecting an appropriate ε for the anonymous reports.

Inverse cumulative distribution Another metric of interest that we estimate is the ICDF. We use it to determine the number of users who take over 10000 steps on a given day, which is $n \cdot Q_{\text{steps}}(1000)$. It appears that the Laplace mechanism maintains an acceptable error (± 2 out of $n = 30$ participants) only for $\varepsilon = 8$ or higher, as shown by fig. 5.5. Since the count depends on the number of participants, adding participants does not improve the error in absolute value. However, the percent error – i.e.,

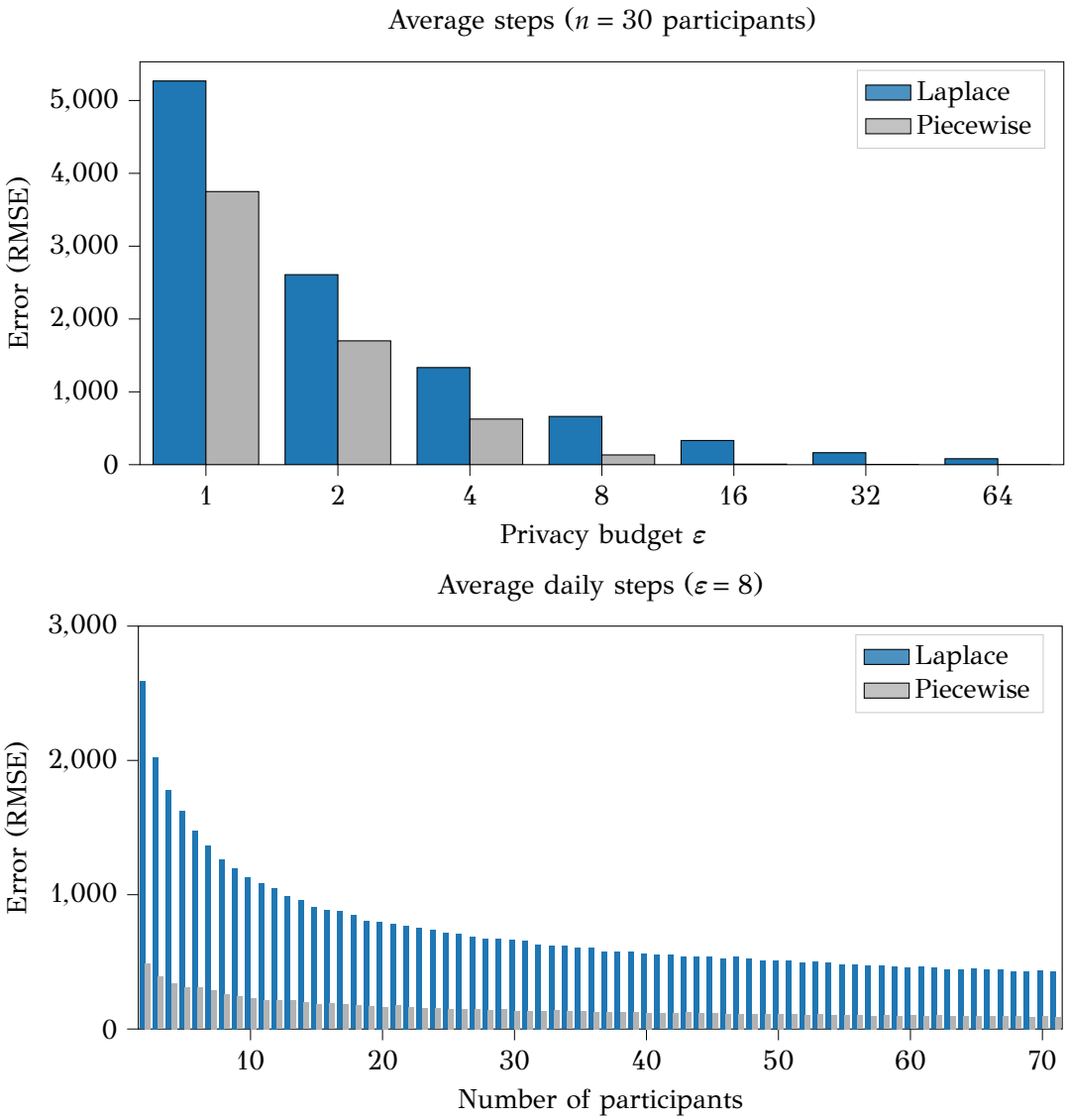


Figure 5.4: RMSE of average step estimates under LDP for varying number of participants n and privacy budget ε. Unsurprisingly, a larger number of participants provides a more accurate estimation of the average. For a same (n, ε) pair, the Piecewise mechanism introduces less noise.

normalized w.r.t. n – decreases with n. This implies that the fraction of participants who met a certain step goal can be estimated with high confidence when the number of participants is large.

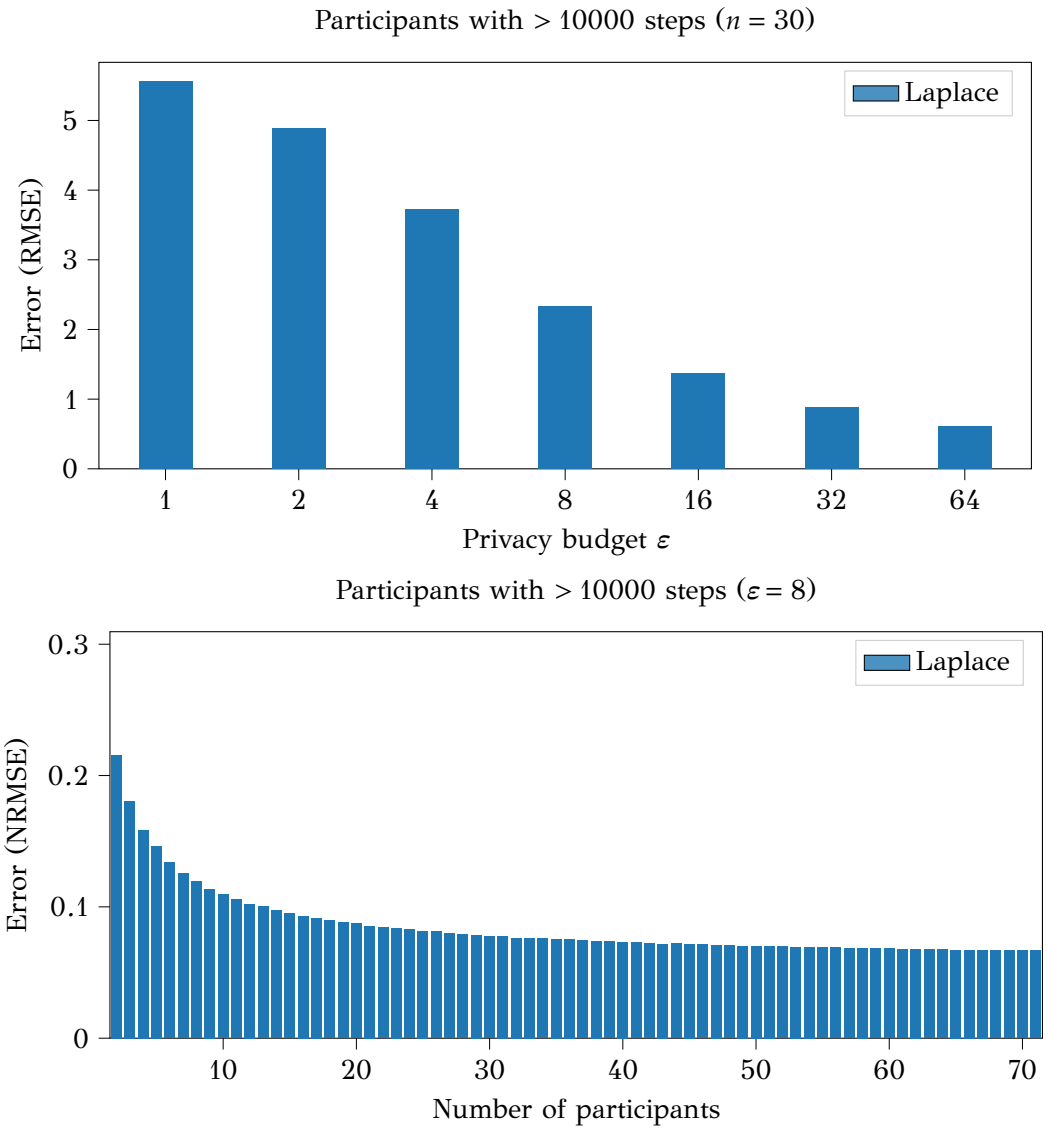


Figure 5.5: RMSE of count estimate ($n \times$ ICDF) for users taking over 10000 steps per day.

Agreement on t-test To measure the agreement on t-tests, we randomly select an even number of participants and split them into two groups of equal size. For instance, the left plot in fig. 5.6 shows the agreement rate for $n = 30$ participants, implying that we randomly selected 30 users and divided them into two groups of 15. As we need to evaluate two types of error, we increased the number of iterations for each Monte Carlo experiment to $n_{\text{trials}} = 1000$. We set the threshold for statistical significance at $a = 0.05$. Running a t-test on two groups of participants in a randomized controlled

trial can either reveal their sample averages to be significantly different ($p < 0.05$) or not ($p \geq 0.05$). The agreement rate indicates the fraction of t-tests conducted on anonymous reports that provide the same result as those on the original records. Our estimator primarily results in type II errors, meaning that a t-test on noisy reports shows no statistical significance where $p < 0.05$ on the original data. Conversely, when the original data does not display a significant difference between the two groups, the agreement rate remains consistently high, regardless of the value of ε. This suggests that this estimation approach is robust against type II errors, as shown in table 5.1. Similar to the sample average case, the Piecewise mechanism offers higher utility than Laplace for the same value of ε. Interestingly, the values of $\varepsilon = 4$ and $\varepsilon = 8$ seem to work well as a threshold between high- and low-utility outcomes, achieving over 90% agreement rate. Therefore, these values of ε may be the ideal choice for practical applications. Remarkably, as shown in the right plot of fig. 5.6, the number of participants appears to have no effect on the results. This is likely due to the t statistic being related to the sample standard deviation of the data, which does not scale with the number of participants since more noisy samples just increase the variance. Figure 5.6 also indicates that the number of random groups with $p < 0.05$ (grey dotted line) is about 70%. This implies that we ran an adequate number of experiments for both cases, $p < 0.05$ and $p \geq 0.05$.

Resilience to linking attacks Evaluating the resilience against linking attacks, the Laplace mechanism seems to provide stronger protection compared to the Piecewise with equal privacy budget. Figure 5.7 shows that for $n = 30$ and $\varepsilon = 8$, Laplace brings the linking rate below 10%. The Piecewise mechanism needs a budget of $\varepsilon = 4$ to achieve the same probability. Overall, the two mechanisms seem to be comparable in terms of privacy-utility tradeoff which can be achieved with different privacy budget. Figure 5.8 depicts such tradeoff for the Laplace mechanism applied to different features. We also stress the fact that in practical attacks, the adversary will not have access to the target's original data, thus the linking probability will be lower. The linking rate obtained in our experiments should be interpreted as a worst-case-scenario result. Another notable observation is that the linking rate decreases with the number of participants, following the similar behavior of the "random guess" curve. This follows the intuition that the needle is harder to find when the haystack is big. In other words, if there is a large number of reports $y_1, \ldots, y_n$, it is likely that a report from another participant will be randomized into a point close to x^* (the original record produced by the target).

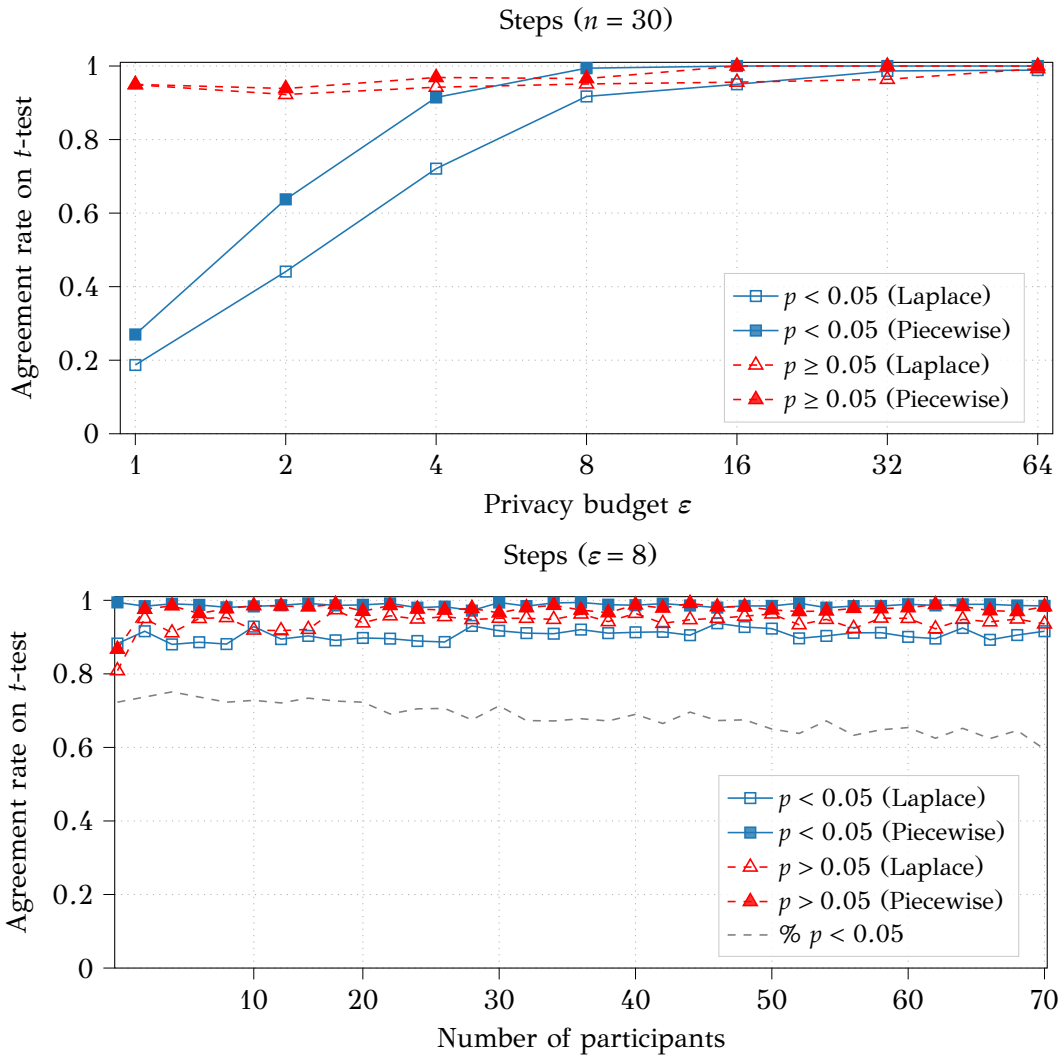


Figure 5.6: Agreement on t-tests for varying privacy budget and number of participants. The agreement rate is divided between the cases where the original data yield statistically significant results ($p < 0.05$) and where they do not ($p \geq 0.05$). A higher agreement rate means more reliability for t-test results under LDP. On the right plot, the grey dotted line indicates the percentage of groups below the p-value threshold ($p < 0.05$).

5.2.3 Limitations and implementation details

In order to give the full picture, we must discuss additional limitations and benefits of LDP with reference to crowdsourcing wearable IoT data.

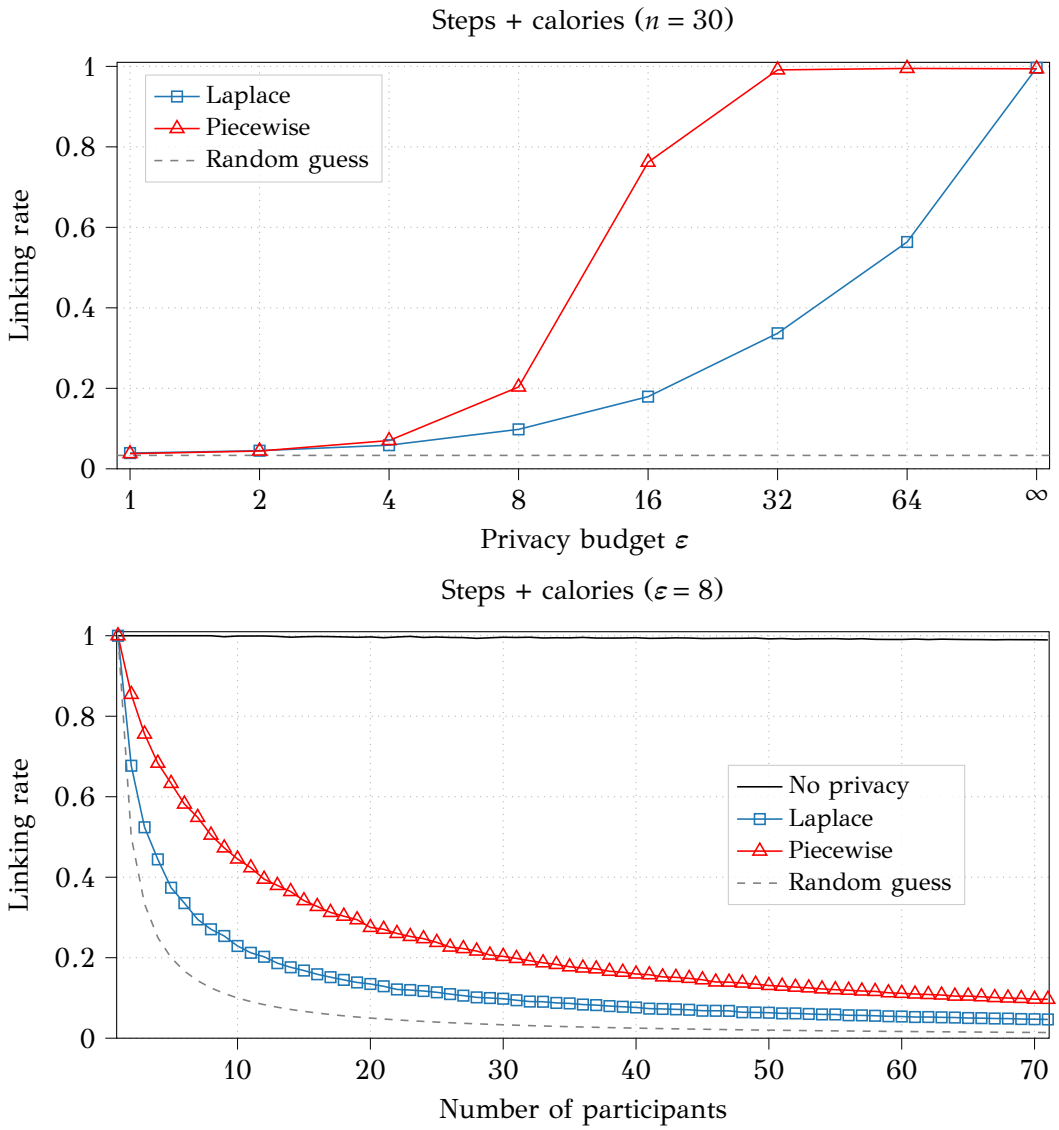


Figure 5.7: Linking rate for varying number of participants and privacy budget ε. A lower linking rate implies more privacy. For a same (n, ε) pair, the Laplace mechanism provide more protection against linking attacks.

Independent reports Under LDP, participants are able to publish multiple independent reports, which means that the analyst has no way of knowing whether two records belong to the same user. This is an intended behavior, which allows for achieving anonymity without further distributing the privacy budget. However, submitting reports independently precludes the possibility of studying them in the temporal dimension, unless the privacy budget is distributed across different records in

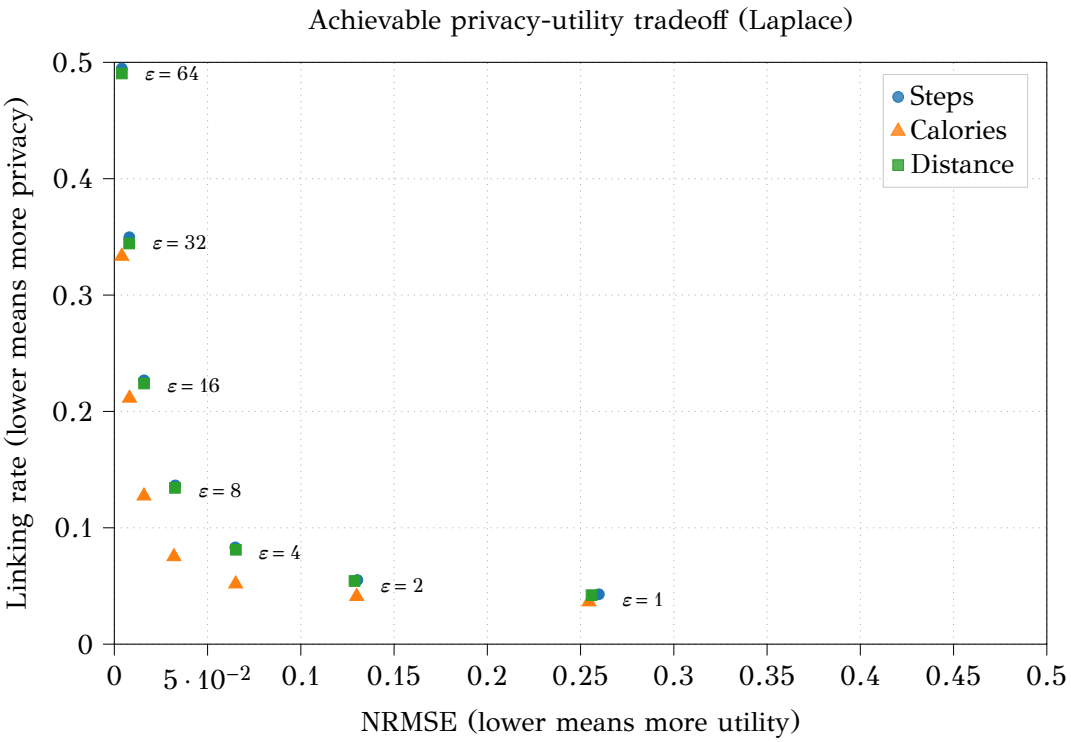


Figure 5.8: Privacy-utility tradeoff achieved by the Laplace mechanism for $n = 30$ participants and different values of ε. It appears that a privacy budget between 4 and 8 offers the best tradeoff.

time, which exponentially increases the noise. Furthermore, while our experiments show that suitably calibrated noise can limit the error, using LDP inevitably reduces data utility and, therefore, the accuracy of the results. Our recommendation is to use LDP in preliminary analyses where limited resources typically reduce the possibility of recruiting many participants in person.

LDP and dataset disclosure LDP provides a way for analysts to share collected data with others, allowing them to reproduce and verify the results. Making crowdsourced data publicly available could greatly benefit the research community and improve the credibility of studies that rely on such information. However, precautions must be taken to prevent accidental privacy breaches, depending on the crowdsourcing framework used. In our crowdsourcing setting, we need to shuffle the data after receiving it to ensure that the published data is not in the same order as it was submitted (encrypted) by the third-party server to the analyst.

Plausible deniability LDP not only helps to prevent linking attacks, but it also offers *plausible deniability* against sensitive inferences. Even if an attacker is able to identify the owner of a randomized record, the information contained within it will still be imprecise. As a result, the adversary will have access to less information than if they had obtained an original record.

LDP in practice We found that using LDP on wearable IoT records provides a higher level of protection than theoretical guarantees suggest. Our results align with other studies on differential privacy, such as those exploring practical membership inference [54] and secrets extraction from machine learning models [16]. These findings highlight the need for additional research into inference attacks against data protected with DP and LDP.

5.2.4 Local differential privacy and anonymity

In our experiments, we have shown empirically that LDP provides a high level of protection against linking attacks, providing anonymity to the participants. In this section, we aim to explain why this happens and what is the relationship between LDP and anonymity.

As mentioned in section 5.2, the dataset collected through the crowdsourcing platform is a sequence of unordered collections of records, where each collection $D[t], t = 1, \ldots, T$ is as follows:

$$D[t] = \{y_1[t], \ldots, y_n[t]\}. \tag{5.22}$$

Since the records are not ordered by participant, an attacker cannot leverage the data as a time series to re-identify her target. The only possible attack involves trying to find identifying information in each record separately. In this scenario, if the actual value of the anonymized record submitted by the target is unknown, the most valuable additional information for the attacker is the original sample. To ensure clarity in our upcoming discussion, we will denote the original record produced by the target participant as $x^* \in \mathbb{R}$, and the collection of anonymized records as $y_1, \ldots, y_n$, omitting the time index t.

Optimal linking criteria In this scenario, the optimal attack to find the target record among $y_1, \ldots, y_n$ can be derived using a maximum a posteriori probability criterion.

$$\hat{y} = \arg \max_{y_i, i=1,\ldots,n} \Pr[X_i = x^* | Y_1 = y_1, \ldots, Y_n = y_n]. \tag{5.23}$$

Since the y_i records are unordered, they have the same prior probability, which means that the maximum a posteriori decision is equivalent to the maximum likelihood decision

$$\hat{y} = \arg \max_{y_i, i=1,\dots,n} \Pr[Y_1 = y_1, \dots, Y_n = y_n | X_i = x^*]. \tag{5.24}$$

Noting that the outcome of $Y_j, j \neq i$ is independent of X_i and Y_i, we can simplify the criterion as follows:

$$\hat{y} = \arg \max_{y_i, i=1,\dots,n} p(y_i|x^*) = \arg \max_{y_i, i=1,\dots,n} \prod_{f=1}^{F} p(y_{i,f}|x_f^*), \tag{5.25}$$

where $p(y_i|x^*)$ denotes the probability density function (PDF) of Y_i given $X_i = x^*$. This PDF is the same for all $i = 1, \dots, n$. The last step is justified by the LDP mechanisms being applied independently to each feature.

If the adopted mechanism is Laplace with privacy budget ε, evaluating the PDF yields

$$\hat{y} = \arg \max_{y_i, i=1,\dots,n} \prod_{f=1}^{F} \exp(-\varepsilon \frac{|y_{i,f} - x_f^*|}{\Delta_f}) \tag{5.26}$$

$$= \arg \max_{y_i, i=1,\dots,n} \sum_{f=1}^{F} -\varepsilon \frac{|y_{i,f} - x_f^*|}{x_{\max,f} - x_{\min,f}} \tag{5.27}$$

$$= \arg \min_{y_i, i=1,\dots,n} \sum_{f=1}^{F} \frac{|y_{i,f} - x_f^*|}{x_{\max,f} - x_{\min,f}}. \tag{5.28}$$

For the Piecewise mechanism, instead,

$$\hat{y} = \arg \max_{y_i, i=1,\dots,n} \prod_{f=1}^{F} \begin{cases} e^{\varepsilon} & \text{if } y_{i,f} \in \left(L(x_f^*), R(x_f^*)\right), \\ 1 & \text{if } y_{i,f} \notin \left(L(x_f^*), R(x_f^*)\right) \end{cases} \tag{5.29}$$

$$= \arg \max_{y_i, i=1,\dots,n} \sum_{f=1}^{F} \chi\left\{y_{i,f} \in \left(L(x_f^*), R(x_f^*)\right)\right\}. \tag{5.30}$$

Bounds on the linking rate Given the optimal attack criteria derived above, we can deduce bounds on the level of protection granted by LDP for each anonymized record. This depends on the privacy budget ε, the number of features F, and the number of participants n. In particular, being $\mathcal{S}$ the success event of a linking attack, the following bounds hold:

- for the Laplace mechanism, letting $\Pr[\mathcal{S}|F, n, \varepsilon, \text{Laplace}] = \gamma$,

$$\gamma \leq 1 - e^{-\varepsilon}\left(1 - \left(1 - \left(\frac{1}{2} - \frac{1}{2}e^{-\frac{2\varepsilon}{F}}\right)^{F}\right)^{n-1}\right); \tag{5.31}$$

- for the Piecewise mechanism, letting $\Pr[\mathcal{S}|F, n, \varepsilon, \text{Piecewise}] = \gamma'$,

$$\gamma' \leq 1 - \frac{e^{\frac{\varepsilon}{3}}}{(e^{\frac{\varepsilon}{3F}} + e^{\frac{\varepsilon}{F}})^{F}}\left(1 - \left(1 - \frac{1}{(e^{\frac{\varepsilon}{3F}} + e^{\frac{\varepsilon}{F}})^{F}}\right)^{n-1}\right). \tag{5.32}$$

Since they are derived by taking into account specific events where a linking attack fails, these bounds are considerably loose. Hence, they should not be considered representative of the level of protection achieved by the corresponding LDP mechanisms, but rather to show such protection exists. Furthermore, they hint that a lower privacy budget and a larger batch of participants limit the linking rate.

The bounds in equations 5.31 and 5.32 can be derived by the probability of failure for the attack $\Pr[\mathcal{S}^c]$. Indeed, this is linked to the success probability by the equation $\Pr[\mathcal{S}] + \Pr[\mathcal{S}^c] = 1$. Thus, finding a lower bound to $\mathcal{S}^c$ means finding an upper bound to $\mathcal{S}$. Such lower bound can be determined by selecting a specific event where the linking attack is guaranteed to fail. We let y_i be the target's anonymous report, while reports y_j, $j \neq i$ are the reports collected from the other users. This implies that, according to eq. 5.20, the attack succeeds if y_i is closer to x^* than any other report. For the Laplace mechanism, with one single feature, the adversary fails if y_i falls outside the region $(x^* - \Delta, x^* + \Delta)$ while at least one other report falls inside such region. If $\Pr[\mathcal{S}^c|1, n, \varepsilon, \text{Lap}] = \eta$

$$\eta \geq \Pr[Y_i \notin (x^* - \Delta, x^* + \Delta) \wedge Y_j \in (x^* - \Delta, x^* + \Delta), j \neq i] \tag{5.33}$$

$$= \Pr[Y_i \notin (x^* - \Delta, x^* + \Delta)]\left(1 - \prod_{j=1, j\neq i}^{n} \Pr[Y_j \notin (x^* - \Delta, x^* + \Delta)]\right) \tag{5.34}$$

$$\geq e^{-\varepsilon}\left(1 - \left(\frac{1}{2} + \frac{1}{2}e^{-2\varepsilon}\right)^{n-1}\right). \tag{5.35}$$

When the reports comprise multiple features, the attack failure is guaranteed if the event of y_i falling outside the region $(x_f^* - \Delta_f, x_f^* + \Delta_f)$ occurs for all features $f = 1, \ldots, F$ (and conversely, all the features of another report fall within the region). Furthermore, each of the F features is randomized with privacy budget ε/F. The bounds, thus,

becomes

$$\Pr[\mathcal{S}^c|F, n, \varepsilon, \text{Laplace}] \geq e^{-\varepsilon}\left(1 - \left(1 - \left(\frac{1}{2} - \frac{1}{2}e^{-\frac{2\varepsilon}{F}}\right)^F\right)^{n-1}\right). \tag{5.36}$$

which leads to eq. 5.31.

A similar reasoning applies to the Piecewise mechanism. In the single-feature case, the adversary fails if y_i falls outside the region $(L(x^*), R(x^*))$ while another report is found inside. Letting $\Pr[\mathcal{S}^c|1, n, \varepsilon, \text{Piecewise}] = \eta'$,

$$\eta' \geq \Pr[Y_i \notin (L(x^*), R(x^*)) \wedge Y_j \in L(x^*), R(x^*)), j \neq i] \tag{5.37}$$

$$= \Pr[Y_i \notin (L(x^*), R(x^*))]\left(1 - \prod_{j=1, j\neq i}^{n} \Pr[Y_j \notin (L(x^*), R(x^*))]\right) \tag{5.38}$$

$$\geq \frac{\tau}{\tau + e^{\varepsilon}}\left(1 - \left(1 - \left(\frac{\tau + e^{\varepsilon} - 1}{\tau + e^{\varepsilon}}\right)^{n-1}\right)\right) \tag{5.39}$$

$$= \frac{e^{\varepsilon/3}}{e^{\varepsilon/3} + e^{\varepsilon}}\left(1 - \left(1 - \left(\frac{e^{\varepsilon/3} + e^{\varepsilon} - 1}{e^{\varepsilon/3} + e^{\varepsilon}}\right)^{n-1}\right)\right) \tag{5.40}$$

Repeating the same reasoning for multiple features, we get

$$\Pr[\mathcal{S}^c|F, n, \varepsilon, \text{Piecewise}] \geq \frac{e^{\frac{\varepsilon}{3}}}{(e^{\frac{\varepsilon}{3F}} + e^{\frac{\varepsilon}{F}})^F}\left(1 - \left(1 - \frac{1}{(e^{\frac{\varepsilon}{3F}} + e^{\frac{\varepsilon}{F}})^F}\right)^{n-1}\right), \tag{5.41}$$

which leads to eq. 5.32.

5.2.5 Alternative applications and limitations

The design of this crowdsourcing platform is well-suited for wearable data, considering how wearable devices are typically used in research. Our platform design ensures that individual users' progress cannot be monitored, only the progress of the entire group. This approach enhances privacy by preventing attackers from accessing individual-level time series data.

In studies focused on health and wearables, users are typically examined as a collective [46], so the inability to conduct individual-level time series analysis does not hinder the usefulness of the gathered data. This design can be applied to other applications with similar characteristics, for example clinical trials that do not rely on wearables but are instead based on self-reported information.

On the other hand, this approach is not appropriate for applications where individual-level information is more important than aggregated data. Recommendation systems for e-commerce and social networks, for example, rely on individual-level time series

or graph information to make accurate predictions [50,51]. In such cases, systems like RAPPOR may achieve a more advantageous privacy-utility tradeoff [30,131].

5.3 Federated Naive Bayes with differential privacy

In this section, we consider a scenario in which an analyst collects data with the purpose of training a machine learning model. In such a scenario, collecting individual data points perturbed with local differential privacy may not be an optimal solution. Our proposed solutions are based on the federated machine learning paradigm, or simply *federated learning*, which involves multiple data owners communicating with a central aggregator (in our case, the analyst) to collaboratively train a machine learning model. In federated learning, the original data never leave the local appliances, and the data owners (which are also called "nodes", using networking terminology) only disclose information in the form of local model updates. This makes more difficult for a "curious" analyst to glean information about individual data points, which is why federated learning gained momentum as a privacy-preserving solution to train machine learning models. Nevertheless, just aggregating information into local updates is not sufficient to guarantee privacy, as we discuss below, and they should also be protected with differential privacy. The difference with the naive solution of simply applying local differential privacy is that perturbing a model update requires to introduce less noise.

Federated learning is already widely established in the context of neural network models, for which several aggregation strategies have already been designed and implemented. A main limitation of neural network models, however, is that they require a large amount of data and adequate hardware equipment for training. Moreover, the volume requirement for the training data increases further when differential privacy guarantees are enforced. This constitutes a barrier for the application of federated learning on wearables, where the supply of data is typically limited. Our contribution in the field of federated learning consists in the design of a federated algorithm to train Naive Bayes models, which are adopted in many machine learning studies that rely on wearable data [42,96,137]. Naive Bayes is a machine learning algorithm that is notorious for its simplicity, which allows to get reliable prediction accuracy with a limited amount of data.

5.3.1 Privacy leaks from machine learning models

While federated learning provides some level of privacy by disclosing aggregated information in the form of local updates, it is not completely immune to privacy leaks. An attacker can still use local updates to infer sensitive information or even recon-

struct the original data. For federated neural networks, an attacker can gain sensitive information by analyzing parameter updates or querying the model. Many works in literature have addressed these issues [135]. For example, Geiping et al. [36] managed to reconstruct images from gradient updates of a federated ResNet model, and Hitaj et al. [47] succeeded in extracting training data by a collaboratively trained generative model. Therefore, when dealing with sensitive training data, federated learning should be used alongside other privacy-preserving techniques, such as differential privacy.

5.3.2 Federated Neural Networks

The term "federated learning" typically refers to federated neural network models, which are the most widely used type of model in the machine learning field. Neural networks are characterized by a large number of trainable parameters, and are able to approximate any kind of function. Although the architecture of these models may vary depending on the application, they are all trained using the stochastic gradient descent (SGD) algorithm, which we discussed in chapter 2. SGD is an iterative procedure that gradually leads the model to converge to a local optimum set of parameters.

The training procedure for neural networks in a federated setting is also iterative. First, the nodes agree on some common initialization for the model parameters θ_0. At each iteration t, nodes perform a certain number of local training steps. Specifically, each node i computes a local update $\theta_{t+1}^{(i)}$ by descending the gradient calculated on its local data. The updates are collected by the central aggregator, which combines them to obtain a global update of the model θ_{t+1}. The most common aggregation approach is simply averaging the received updates $\theta_{t+1}^{(i)}$ [85], but other methods have been proposed [73, 107, 117].

In order to train federated neural network models with differential privacy, the standard approach is to clip and perturb the gradient with noise during the gradient descent process. This solution was proven to be highly effective to prevent unwanted memorization of training data in the model [16].

Despite their popularity, federated neural network models have many drawbacks. One is that they require a large sample of data to achieve high prediction performance in the final model. Furthermore, neural networks are mainly used as black box models, making it hard to determine the rationale behind a prediction.

In health applications, interpretable models that can be understood and adapted by experts are arguably preferable. Additionally, while crowdsourcing wearable data may allow access to a wide pool of records, it might be difficult to gather a sufficient amount of data points to train state-of-the-art neural network models.

5.3.3 Federated Naive Bayes

One main contribution of our research in the field of federated machine learning is the design of an algorithm to train Naive Bayes models in a federated setting under differential privacy guarantees [38, 79]. Albeit simple in its design, Naive Bayes has proven to be an effective machine learning algorithm for classification. In Naive Bayes models, different features contribute independently to the prediction. While this may be viewed as a limitation, it also renders the algorithm virtually immune to overfitting, as shown by experimental results [138]. Moreover, Naive Bayes has shown effectiveness even when the assumption of independence between features does not hold [109]. The original algorithm, outlined in chapter 2, leverages basic statistics computed on the training data (mainly counts and sums) to produce a classification model. In our federated design, n_{nodes} data owners, also called *nodes*, run these queries locally on their own data and randomize the output to achieve differential privacy. Then, a central aggregator collects the resulting noisy parameters and aggregates them to produce the final model.

Our novel algorithm aims to facilitate the training of simple models in a federated setting. Although federated learning is widely explored in academic research, training federated models is often challenging due to limited data availability or resource constraints, as we will further explain in the remainder of this section. Introducing Federated Naive Bayes, we present a class of models that can be trained through a single exchange between the nodes and the central aggregator. The advantage of completing training in a single exchange is twofold. Firstly, it reduces resource requirements, making it more feasible in resource-constrained scenarios. Secondly, when applying differential privacy, this approach offers a favorable privacy-utility tradeoff as the privacy budget does not need to be divided across multiple iterations. This latter benefit is also confirmed by our experimental results.

Previous work The original centralized algorithm that applies differential privacy to Naive Bayes was proposed by Vaidya et al. [127]. Our work extends this algorithm to make it applicable to a federated setting. Additionally, prior works have proposed solutions to train Naive Bayes models on partitioned data. In [58], the authors designed an algorithm to train Naive Bayes on horizontally partitioned data, while [126] proposed solution for vertically partitioned data. Both solutions, however, rely on cryptographic methods that protect the data during the training procedure, but do not provide guarantees that information is not leaked from the resulting model. Another method to protect horizontal data partitions during training relies on semi-trusted mixers [134]. An algorithm that combines homomorphic encryption and differential privacy was proposed in [74]. However, the algorithm estimates conditional probabil-

ities of numerical features using histograms, rather than with the standard Gaussian Naive Bayes approach. Furthermore, the paper does not report how the accuracy is affected by the privacy budget and distribution of the data among the nodes. In [53], the authors studied how differential privacy affects the accuracy of Naive Bayes, in a federated setting where data are vertically partitioned. The contribution of our work can be considered complementary, since we consider a federated setting where data are horizontally partitioned.

5.3.4 Algorithm design

In order to train a federated Naive Bayes model, the central aggregator should be able to collect information that allows to compute the prior and conditional probabilities for each feature and class. In order to guarantee differential privacy, all the query results must be disclosed after being adequately randomized. We employed the Laplace mechanism to perturb the output of each query. This choice was made because the Laplace mechanism is a commonly used approach for handling numerical queries and it scales effectively with the number of data points within each partition [29]. The parameter ε' of the Laplace mechanism is decided according to the privacy budget ε and to the number of queries asked to each node[1].

Mirroring equation 2.3, prior probabilities $p_Y(y), y = 1, \dots, C$ can be simply estimated by collecting from each node $D^{(j)}, j = 1, \dots, n_{\text{nodes}}$ the number of samples per each class $n_1^{(j)}, \dots, n_C^{(j)}$ and by computing

$$p_Y(y) = \frac{\sum_{j=1}^{n_{\text{nodes}}} n_y^{(j)}}{\sum_{j=1}^{n_{\text{nodes}}} \sum_{y'=1}^{C} n_y^{(j)}}. \tag{5.42}$$

Similarly, conditional probabilities for categorical features are estimated according to

$$p_{X_f|Y}(x_f|y) = \frac{\sum_{j=1}^{n_{\text{nodes}}} m_{x_f y}^{(i)}}{\sum_{j=1}^{n_{\text{nodes}}} n_y^{(j)}} \tag{5.43}$$

where $m_{x_f y}^{(i)}$ is the number of samples with feature $f \in \mathcal{F}_{\text{cat}}$ equal to v for node $D^{(i)}$. For numerical features, Gaussian Naive Bayes requires to compute the parameters μ_{fy}, σ_{fy}^2 of the assumed normal distribution of each feature for each class. The mean μ_{fy} can

[1] In principle, each node may have a different privacy budget, but herein we assume there is a common value of ε for all the nodes.

be derived by querying the sample sum

$$S_{fy}^{(j)} = \sum_{x \text{ in class } y} x_f^{(j)} \tag{5.44}$$

for each numerical feature $f \in \mathcal{F}_{\text{num}}$ and class y from each node $D^{(i)}$ as

$$\mu_{fy} = \frac{\sum_{j=1}^{n_{\text{nodes}}} S_{fy}^{(j)}}{\sum_{j=1}^{n_{\text{nodes}}} n_y^{(j)}}. \tag{5.45}$$

For the variance σ_{fy}^2, a straightforward solution would be to calculate it by having each node computing the sum of squared deviations from the sample average μ_{fy}, i.e.,

$$\sigma_{fy}^2 = \sum_{x_f \text{ in class } y} (x_f^{(j)} - \mu_{fy})^2. \tag{5.46}$$

However, this solution would require two exchanges between each node and the central aggregator, since μ_{fy} would need to be sent back to the nodes to compute the squared deviations. A solution that enables the central aggregator to compute means and variances with a single exchange leverages the relation between second moment, mean, and variance, i.e., $\sigma_{fy}^2 = s_{fy} - \mu_{fy}^2$. Therefore, in our algorithm the nodes are asked to compute the sum of the squared samples

$$Q_{fy}^{(j)} = \sum_{x_f \text{ in class } y} \left(x_f^{(j)}\right)^2 \tag{5.47}$$

and the variance is obtained by the centralized aggregator as

$$\sigma_{fy}^2 = \frac{\sum_{i=1}^{n_{\text{nodes}}} Q_{fy}^{(j)}}{\sum_{j=1}^{n_{\text{nodes}}} n_y^{(j)}} - \mu_{fy}^2. \tag{5.48}$$

Enforcing differential privacy In order to ensure that each partition $D^{(j)}$ is ε-differentially private, all queries must be secured with the Laplace mechanism $L_{\varepsilon'}$. Since queries on different classes are computed on disjoint subsets of $D^{(i)}$ (a sample cannot belong to multiple classes at the same time), from a privacy perspective they count as a single query. On the other hand, queries on different features of $D^{(i)}$ are counted separately, since they involve the same entries. Therefore each node overall is asked:

- 1 class counting query, i.e, $n_y^{(i)}$;

Algorithm 1 Local queries computed by j-th node

Require: Data partition $D^{(j)}$, privacy budget ε

for all classes $y = 1, \ldots, C$ **do**
  $\varepsilon' \leftarrow \dfrac{\varepsilon}{1 + F_{\text{cat}} + 2F_{\text{num}}}$
  $\tilde{n}_y^{(j)} \leftarrow \textsc{Laplace}(|\{x^{(j)} : x^{(j)} \text{ in class } y\}|, \varepsilon')$
  for all categorical features $f \in \mathcal{F}_{\text{cat}}$ **do**
    for all categories x_f of feature f **do**
      $\tilde{m}_{x_f y}^{(i)} \leftarrow \textsc{Laplace}(|\{x^{(j)} : x^{(j)} \text{ in class } y \text{ and } x_f^{(j)} = x_f\}|, \varepsilon')$
    end for
  end for
  for all numerical features $f \in \mathcal{F}_{\text{num}}$ **do**
    $\tilde{S}_{fy}^{(j)} \leftarrow \textsc{Laplace}(\sum_{x_f \text{ in class } y} x_f^{(j)}, \varepsilon')$
    $\tilde{Q}_{fy}^{(i)} \leftarrow \textsc{Laplace}(\sum_{x_f \text{ in class } y} (x_f^{(j)})^2, \varepsilon')$
  end for
end for
return $\tilde{n}^{(j)}, \tilde{m}^{(j)}, \tilde{S}^{(j)}, \tilde{Q}^{(j)}$

Algorithm 2 Centralized aggregation

Require: $\tilde{n}^{(j)}, \tilde{m}^{(j)}, \tilde{S}^{(j)}, \tilde{Q}^{(j)}$ for $j = 1, \ldots, n_{\text{nodes}}$

for all classes $y = 1, \ldots, C$ **do**
  $\tilde{p}_Y(y) \leftarrow \dfrac{\sum_{i=1}^{n_{\text{nodes}}} \tilde{n}_y^{(j)}}{\sum_{j=1}^{n_{\text{nodes}}} \sum_{y'=1}^{C} \tilde{n}_y^{(j)}}$
  for all categorical features $f \in \mathcal{F}_{\text{cat}}$ **do**
    for all categories x_f of f **do**
      $\tilde{p}_{X_f|Y}(x_f|y) \leftarrow \dfrac{\sum_{j=1}^{n_{\text{nodes}}} \tilde{m}_{x_f y}^{(i)}}{\sum_{j=1}^{n_{\text{nodes}}} \tilde{n}_y^{(j)}}$
    end for
  end for
  for all numerical features $f \in \mathcal{F}_{\text{num}}$ **do**
    $\tilde{\mu}_{fy} \leftarrow \dfrac{\sum_{j=1}^{n} \tilde{S}_{fy}^{(i)}}{\sum_{i=1}^{n_{\text{nodes}}} \tilde{n}_y^{(j)}}$
    $\tilde{\sigma}_{fy}^2 \leftarrow \dfrac{\sum_{j=1}^{n_{\text{nodes}}} \tilde{Q}_{fy}^{(j)}}{\sum_{j=1}^{n_{\text{nodes}}} \tilde{n}_y^{(j)}} - \tilde{\mu}_{fy,}^2$
  end for
end for
return $\tilde{p}_Y, \tilde{p}_{X_f|Y}, \tilde{\mu}, \tilde{\sigma}^2$

- F_{cat} category counting queries, i.e., $m^{(i)}_{x_f y}, f \in \mathcal{F}_{\text{cat}}$;
- $2F_{\text{num}}$ sum queries on numerical features, i.e., $S^{(i)}_{fy}, Q^{(i)}_{fy}, f \in \mathcal{F}_{\text{num}}$.

Therefore, in order to guarantee a privacy budget of ε, the parameter ε' of the Laplace mechanism to be applied to each query is given by

$$\varepsilon' = \frac{\varepsilon}{1 + F_{\text{cat}} + 2F_{\text{num}}}, \tag{5.49}$$

in congruence with the centralized ε-DP Naive Bayes by Vaidya et al. [127].

Training procedure As mentioned above, the training procedure consists of a single exchange for each node with the central aggregator. Each node i computes locally the counting and sum queries on its own partition according to algorithm 1. The query results $\tilde{n}^{(j)} \in \mathbb{Z}^C, \tilde{m}^{(j)} \in \mathbb{Z}^{F_{\text{cat}} \times C}, \tilde{S}^{(j)} \in \mathbb{R}^{F_{\text{num}} \times C}, \tilde{Q}^{(j)} \in \mathbb{R}^{F_{\text{num}} \times C}$ are sent to the central aggregator, which collects them across all nodes and computes the overall parameters of the model according to algorithm 2. Finally, the computed parameters are used to infer the class of a sample, as per the original Naive Bayes algorithm:

$$\hat{y} = \arg \max_{y=1,\dots,C} p_Y(y) \prod_{f=1}^{F} p_{X_f|Y}(x_f|y). \tag{5.50}$$

5.3.5 Extensions

Online updates Algorithm 2 takes as input the query results from all nodes. In reality, we expect the training procedure to be executed in an asynchronous fashion, with each participant sending its results at a different time and the aggregator performing subsequent updates to the Naive Bayes parameters. This also allows new nodes to join the overlay network and collaborate in training the model. One straightforward way to make this possible is to store the responses of each node, and recompute the model parameters every time a new set of query responses is sent by a node. However, this is inefficient and requires the central aggregator to store unnecessary information.

An alternative consists in storing the following aggregated information:

- the total number $\tilde{\nu}_y$ of samples for each class;
- the total counters of categorical features $\tilde{M}_{fvy}$ for each class;
- the values of mean and variance of all the numerical features $\tilde{\mu}_{fy}, \tilde{\sigma}^2_{fy}$ for each class.

Notice that the parameters $\tilde{p}_Y(y)$ and $\tilde{p}_{X_f|Y}(x_f|y)$ can be computed from $\tilde{\nu}_y$ and $\tilde{M}_{x_f y}$ as

$$\tilde{p}_Y(y) = \frac{\tilde{\nu}_y}{\sum_{y'=1}^{C} \tilde{\nu}_{y'}}, \quad \tilde{p}_{X|Y}(x_f|y) = \frac{\tilde{M}_{x_f y}}{\tilde{\nu}_y}, \tag{5.51}$$

so the stored variables completely characterize the model.

5.3.6 Experimental evaluation

We evaluated our approach using six standard datasets obtained from the UCI repository[2]. The statistics of each datasets are listed in table 5.3. For datasets without a standardized train/test split, we performed a 90/10 split using a fixed random seed throughout our experiments.

We implement three variants of Naive Bayes:

1. a "vanilla" Naive Bayes model with access to the entire centralized dataset and no differential privacy protection;
2. a differentially-private centralized Naive Bayes based on [127];
3. our differentially-private federated Naive Bayes based on algorithms 1 and 2.

We conducted a Monte Carlo analysis to test each variant of our approach on all datasets. For the differentially-private variants, we used values of ε between 0.01 and 10. For the federated approach, we tested with various numbers of data partitions (nodes) $n_{\text{nodes}} \in [1, 10, 100, 1000]$ whenever possible. However, we skipped the higher values of n on small datasets, making sure that each partition had at least two data points. We repeated each experiment 1000 times to account for the randomness in both the data partitioning across federated entities and the sampling of Laplace noise. Unless otherwise stated, we report the mean of those 1000 trials. All the code and data used in this work are available on GitHub[3].

Figure 5.9 plots the accuracy of the differentially-private Naive Bayes models as a function of the privacy budget ε, including both the centralized approach, and the federated approach with different number of data partitions n. The score obtained by a non-differentially-private model is shown as an upper bound of attainable performance.

On most datasets, both the centralized and federated models exhibit a similar pattern of performance improvement as the privacy budget increases. The centralized model typically follows an S-shaped curve, starting with performance similar to random guessing for very low values of ε, but quickly improving towards the baseline

[2] https://archive.ics.uci.edu/ml/datasets.php

[3] `https://github.com/thomasmarchioro3/FederatedNaiveBayesDP`

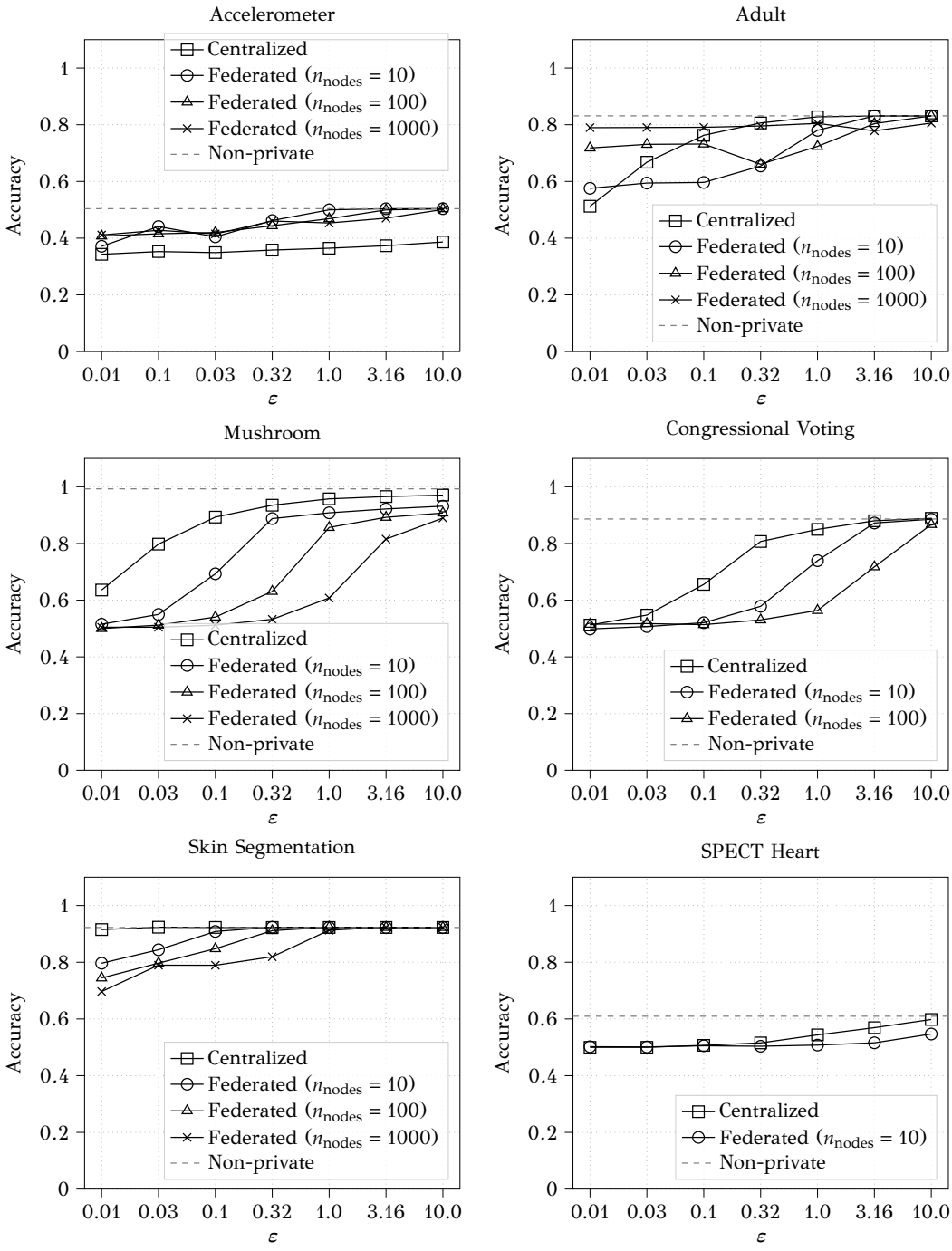


Figure 5.9: Performance analysis (accuracy versus privacy budget) of our proposed federated Naive Bayes algorithm. Our solution is compared with the non-private Naive Bayes and with the centralized differentially private algorithm by Vaidya et al. [127].

Dataset	Samples	Labels	F_{num}	F_{cat}	Predefined train/test split
Accelerometer	153,000	3	3	0	no
Adult	48,842	2	6	8	yes (2:1)
Congressional Voting	435	2	0	16	no
Mushroom	8,124	2	0	22	no
Skin Segmentation	245,057	2	3	0	no
SPECT Heart	267	2	0	22	yes (3:7)

Table 5.3: Datasets used in the evaluation of federated Naive Bayes with differential privacy.

accuracy. The federated model also follows this curve but with a delay proportional to the number of partitions. This behavior is especially evident in the Mushroom and Congressional Voting datasets. However, in the Mushroom dataset, it is difficult to fully observe the left-hand tail of the curve with $\varepsilon = 10^{-2}$. The Skin Segmentation dataset, on the other hand, has a large sample count, resulting in the right-hand tail of the S-shaped curve. For this dataset, the federated models only slightly dip below the maximum performance for very low values of ε. The performance curve of the federated model on the SPECT Heart dataset shows a prolonged decline on the left-hand tail. This is likely due to the limited size of the dataset and possibly to the presence of non-independent features. It is worth noting that even the accuracy of the centralized and non-private Naive Bayes algorithm on this dataset does not surpass random guessing by a significant margin. This suggest that the original algorithm itself is not suitable for this specific dataset.

These findings emphasize that the value of ε at which the model is severely impacted varies significantly depending on the dataset. Therefore, it is crucial to assess this threshold on a domain-specific basis. Our results also show how switching to a federated setting, and thus avoiding the data collection phase, only requires a small increase to the privacy budget in order to achieve the same accuracy, with the possibility of even surpassing a centralized solution in certain limited scenarios.

5.3.7 Alternative applications and limitations

A main limitation in our assessment of federated Naive Bayes is the lack of available data for machine learning applications based on wearables. Differently from comparative studies, machine learning applications require to use labeled datasets to train and test machine learning model. For example, if we want to demonstrate that federated Naive Bayes can be used to predict stress based on wearable data, as done in [96], we would need a dataset containing pairs of wearable records and stress levels.

Nevertheless, despite this limitation, the consistent performance observed on various benchmark datasets indicates that federated Naive Bayes can be successfully utilized for a wide range of applications while maintaining a favorable balance between privacy and utility. Moreover, this suggests that federated Naive Bayes can be applied to any type of dataset for which the original algorithm is suitable, making our contribution highly versatile and applicable in different scenarios [10, 71, 104].

5.4 Takeaways

- Decentralized solutions can be used to anonymize the data directly at the user's side, without the need to rely on a trusted data curator.
- Enforcing privacy guarantees in an online crowdsourcing setting can provide researchers with access to a wider pool of data, while also protecting individuals who contribute their data.
- Decentralized solutions based on differential privacy can be applied to two main use-cases in the context of wearable data: (i) comparative studies based on wearables; (ii) federated machine learning for wearable data applications.
- Users can protect their wearable records with local differential privacy (LDP), perturbing them with properly calibrated noise. A privacy budget below 8 offers a high level of protection against re-identification threats.
- After collecting records protected with LDP, an analyst can still compute useful metrics such as average values, cumulative distribution, and p-value for statistical tests, which are used in comparative studies. All these metrics can be evaluated within a reasonable error margin by recruiting at least 30 participants and using a privacy budget between 4 and 8.
- In federated learning, many data partitions (also called "nodes") exchange updates with a central aggregator to collaboratively train a machine learning model.
- Federated neural network models require a large amount of data and are not very interpretable, making them less suitable for health applications based on wearable data. Furthermore, the training procedure needs several exchanges between the nodes and the central aggregator. Conversely, simpler federated models are more interpretable and require less data.
- The Federated Naive Bayes algorithm proposed in this thesis allows to train Naive Bayes models with differential privacy guarantees. The algorithm requires a single exchange between the data owners and the central aggregator.

- Overall, Federated Naive Bayes with differential privacy offers comparable performance to its centralized counterpart. This was demonstrated by testing it on 6 benchmark datasets.

Chapter 6
Conclusion

We conclude this thesis by summarizing the main results obtained in our works. Throughout our study, we have explored and addressed various key aspects of wearable data privacy.

Our research reveals that time series of data recorded by wearables, such as steps taken and calories burned, can become fingerprints for the device users and re-identify them in anonymized datasets. This implies that enforcing k-anonymity on personal information is not sufficient to protect datasets of wearable records.

To protect these datasets, we propose solutions that can be adopted by data publishers in a centralized setting and by device users in a decentralized setting. Decentralized solutions are generally preferable since they do not require entrusting the data to a curator who is not the device owner. However, in many cases, centralized approaches are necessary, such as when data are collected directly by an organization that purchased devices and recruited participants. For such cases, we developed guidelines that data publishers can use to sanitize the data before publication.

We also cover a decentralized setting created by an online crowdsourcing platform, which can serve as a meeting point for analysts and wearable device users. We developed privacy-preserving solutions that enable users to submit data on this platform while maintaining anonymity under theoretical guarantees. Our proposed solutions are based on differential privacy and involve protecting the data with noise before submitting it. We also studied the usability of these solutions by considering the comparative studies and machine learning applications mentioned above.

Overall, this thesis demonstrates that decentralized solutions with proper privacy guarantees can be feasibly adopted while maintaining adequate levels of utility. Our hope is for this work to contribute to the adoption of decentralized solutions in real-world applications of data from wearable devices in health research.

6.1 Synopsis of contributions

The contributions of this thesis in relation to its research questions can be summarized as follows:

***RQ1**: What are the practical risks of participating in a health study that makes use of wearables? To what extent do these risks impact the privacy of a user?*

To address this question, we studied existing public datasets of wearable records and found that most of them do not satisfy the fundamental k-anonymity requirements for demographic and physical characteristics, which can expose participants to re-identification threats. Additionally, we demonstrated that even if k-anonymity is enforced on these characteristics, an attacker may still be able to recognize a target by (i) linking wearable records of anonymous participants with additional records belonging to their target, or (ii) inferring personal characteristics of the participants based on their wearable records.

***RQ2**: What can data collectors/publishers do to protect wearable data before disclosing them?*

We thoroughly studied existing solutions adopted to protect datasets of wearable records. Combining these with the insights derived from our studies, we developed a set of guidelines to mitigate re-identification threats against these datasets. These guidelines include the enforcement of k-anonymity, aggregation of personal information, resampling, and quantization. We applied these metrics to LifeSnaps, a dataset of Fitbit records collected from 71 participants.

***RQ3**: What can device users do in order to protect their data before disclosing them? Are there viable decentralized solutions to anonymize/sanitize wearable data while preserving their utility for health researchers?*

We evaluated decentralized privacy-preserving solutions that can be applied directly by device users on their local appliances. These solutions are based on the application of noise to provide differential privacy guarantees. We adapted and extended existing differentially private mechanisms to anonymize users' data. During this process, we investigated two main applications: (i) comparative studies based on wearable data and (ii) federated learning with wearables. In the first use-case, we extensively evaluated the usability of records protected with local differential privacy (LDP). Additionally, we devised a design for an online crowdsourcing platform that can be used to collect multiple records with a fixed privacy budget, without requiring additional noise. For machine learning use-cases, we considered the usability of federated learning for wearable data. As federated neural networks are complex and lack interpretability, we designed a federated

version of the Naive Bayes algorithm, which is more suitable for interpretable health-related applications. We evaluated the performance of federated Naive Bayes with differential privacy on benchmark datasets and demonstrated that it achieves comparable accuracy to its centralized counterpart.

6.1.1 Future work

Although we put forth our best effort in addressing the research questions above, there are some points that we were unable to cover in our research, which may inspire future work:

- In our publications, we mainly focused on re-identification attacks based on steps, calories, and distance records, adding heart rate in some cases. These are measurements that most users share on fitness social networks such as the Fitbit community. However, modern consumer-level wearables track more parameters such as respiratory rate, blood pressure, and glucose levels [112]. It would be interesting to investigate how these measurements contribute to re-identification.
- The set of guidelines that we devised to protect wearable data are sensible solutions to mitigate de-anonymization threats. However, these guidelines should be quantitatively evaluated by measuring their effectiveness against different re-identification attacks.
- As mentioned in chapter 5, a main limitation of our assessment of federated Naive Bayes is the lack of available datasets. We tested our algorithm on six benchmark datasets, which suggest that it can achieve a favorable privacy-utility tradeoff in most applications. However, its applicability to wearable data needs to be properly evaluated.

Bibliography

[1] Mohamed Alloghani, Mohammed M Alani, Dhiya Al-Jumeily, Thar Baker, Jamila Mustafina, Abir Hussain, and Ahmed J Aljaaf. A systematic review on the status and progress of homomorphic encryption technologies. *Journal of Information Security and Applications*, 48:102362, 2019.

[2] Abdulmajeed Alqhatani and Heather Richter Lipford. "There is nothing that I need to keep secret": Sharing Practices and Concerns of Wearable Fitness Data. In *SOUPS@ USENIX Security Symposium*, 2019.

[3] Matar Abdullah Alzahrani, Louise Ada, and Catherine M Dean. Duration of physical activity is normal but frequency is reduced after stroke: an observational study. *Journal of physiotherapy*, 57(1):47–51, 2011.

[4] Karuna Arava and Sumalatha Lingamgunta. Adaptive k-anonymity approach for privacy preserving in cloud. *Arabian Journal for Science and Engineering*, 45(4):2425–2432, 2020.

[5] H Ceren Ates, Ali K Yetisen, Firat Güder, and Can Dincer. Wearable devices for the detection of COVID-19. *Nature Electronics*, 4(1):13–14, 2021.

[6] Brian A'hearn, Franco Peracchi, and Giovanni Vecchi. Height and the normal distribution: Evidence from italian military data. *Demography*, 46(1):1–25, 2009.

[7] Yang Bai, Ryan Burns, Nancy Gell, and Wonwoo Byun. A randomized trial to promote physical activity in adult pre-hypertensive and hypertensive patients. *Journal of Sports Sciences*, 40(14):1648–1657, 2022.

[8] Sándor Beniczky, Philippa Karoly, Ewan Nurse, Philippe Ryvlin, and Mark Cook. Machine learning and wearable devices of the future. *Epilepsia*, 62:S116–S124, 2021.

[9] Daniel Bernau, Jonas Robl, and Florian Kerschbaum. Assessing differentially private variational autoencoders under membership inference. *arXiv preprint arXiv:2204.07877*, 2022.

[10] Daniel Berrar. Bayes' theorem and naive bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 403:412, 2018.

[11] Asia J Biega, Peter Potash, Hal Daumé, Fernando Diaz, and Michèle Finck. Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 399–408, 2020.

[12] C Alan Boneau. The effects of violations of assumptions underlying the t test. *Psychological bulletin*, 57(1):49, 1960.

[13] Antoine Boutet, Carole Frindel, Sébastien Gambs, Théo Jourdan, and Rosin Claude Ngueveu. Dysan: Dynamically sanitizing motion sensor data against sensitive inferences through adversarial networks. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 672–686, 2021.

[14] Wlodzimierz Bryc. *The normal distribution: characterizations with applications*, volume 100. Springer Science & Business Media, 2012.

[15] Nancy F Butte, Ulf Ekelund, and Klaas R Westerterp. Assessing physical activity using wearable monitors: measures of physical activity. *Med Sci Sports Exerc*, 44(1 Suppl 1):S5–12, 2012.

[16] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, volume 267, 2019.

[17] Tony F Chan, Gene H Golub, and Randall J LeVeque. Updating formulae and a pairwise algorithm for computing sample variances. In *COMPSTAT 1982 5th Symposium held at Toulouse 1982: Part I: Proceedings in Computational Statistics*, pages 30–41. Springer, 1982.

[18] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.

[19] Man Lai Cheung, Ka Yin Chau, Michael Huen Sum Lam, Gary Tse, Ka Yan Ho, Stuart W Flint, David R Broom, Ejoe Kar Ho Tso, and Ka Yiu Lee. Examining consumers' adoption of wearable healthcare technology: The role of health attributes. *International journal of environmental research and public health*, 16(13):2257, 2019.

[20] Michelle M Christovich. Why should we care what fitbit shares? A proposed statutory solution to protect sensitive personal fitness information. *Hastings Comm. & Ent. LJ*, 38:91, 2016.

[21] Victor Costan and Srinivas Devadas. Intel sgx explained. *Cryptology ePrint Archive*, 2016.

[22] Hila Ariela Dafny, Stephanie Champion, Lemlem G Gebremichael, Vincent Pearson, Jeroen M Hendriks, Robyn A Clark, Maria Alejandra Pinero de Plaza, Aarti Gulyani, Sonia Hines, Alline Beleigoli, et al. Cardiac rehabilitation, physical activity, and the effectiveness of activity monitoring devices on cardiovascular patients: An umbrella review of systematic reviews. *European Heart Journal-Quality of Care and Clinical Outcomes*, page qcad005, 2023.

[23] Alexia Dini Kounoudes, Georgia M Kapitsaki, and Ioannis Katakis. Enhancing user awareness on inferences obtained from fitness trackers data. *User Modeling and User-Adapted Interaction*, pages 1–48, 2023.

[24] Josep Domingo-Ferrer and Vicenç Torra. A critique of k-anonymity and some of its enhancements. In *2008 Third International Conference on Availability, Reliability and Security*, pages 990–993. IEEE, 2008.

[25] Yujie Dong, Adam Hoover, Jenna Scisco, and Eric Muth. A new method for measuring meal intake in humans via automated wrist motion tracking. *Applied psychophysiology and biofeedback*, 37(3):205–215, 2012.

[26] Ashutosh Dhar Dwivedi, Gautam Srivastava, Shalini Dhar, and Rajani Singh. A decentralized privacy-preserving healthcare blockchain for iot. *Sensors*, 19(2):326, 2019.

[27] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006.

[28] Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5*, pages 1–19. Springer, 2008.

[29] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[30] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[31] Lynne M Feehan, Jasmina Geldman, Eric C Sayre, Chance Park, Allison M Ezzat, Ju Young Yoo, Clayon B Hamilton, and Linda C Li. Accuracy of fitbit devices: systematic review and narrative syntheses of quantitative data. *JMIR mHealth and uHealth*, 6(8):e10527, 2018.

[32] Ty Ferguson, Timothy Olds, Rachel Curtis, Henry Blake, Alyson J Crozier, Kylie Dankiw, Dorothea Dumuid, Daiki Kasai, Edward O'Connor, Rosa Virgara, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. *The Lancet Digital Health*, 4(8):e615–e626, 2022.

[33] Daniel Fuller, Emily Colwell, Jonathan Low, Kassia Orychock, Melissa Ann Tobin, Bo Simango, Richard Buote, Desiree Van Heerden, Hui Luan, Kimberley Cullen, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. *JMIR mHealth and uHealth*, 8(9):e18694, 2020.

[34] Robert Furberg, Julia Brinton, Michael Keating, and Alexa Ortiz. Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016, May 2016.

[35] Simson Garfinkel, John M Abowd, and Christian Martindale. Understanding database reconstruction attacks on public data. *Communications of the ACM*, 62(3):46–53, 2019.

[36] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

[37] Lodovico Giaretta, Thomas Marchioro, Evangelos Markatos, and Šarūnas Girdzijauskas. Towards a decentralized infrastructure for data marketplaces: narrowing the gap between academia and industry. In *Proceedings of the 1st International Workshop on Data Economy*, pages 49–56, 2022.

[38] Lodovico Giaretta, Thomas Marchioro, Evangelos Markatos, and Sarunas Girdzijauskas. Towards a realistic decentralized naive bayes with differential privacy. 2023.

[39] Lodovico Giaretta, Ioannis Savvidis, Thomas Marchioro, Šarūnas Girdzijauskas, George Pallis, Marios D Dikaiakos, and Evangelos Markatos. Pds 2: A user-centered decentralized marketplace for privacy preserving data processing. In *2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW)*, pages 92–99. IEEE, 2021.

[40] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[41] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[42] Laure Gossec, Frédéric Guyard, Didier Leroy, Thomas Lafargue, Michel Seiler, Charlotte Jacquemin, Anna Molto, Jérémie Sellam, Violaine Foltz, Frederique Gandjbakhch, et al. Detection of flares by decrease in physical activity, collected using wearable activity trackers in rheumatoid arthritis or axial spondyloarthritis: an application of machine learning analyses in rheumatology. *Arthritis care & research*, 71(10):1336–1343, 2019.

[43] Sanna Hakala, Heikki Kivistö, Teemu Paajanen, Annaliisa Kankainen, Marjo-Riitta Anttila, Ari Heinonen, and Tuulikki Sjögren. Effectiveness of distance technology in promoting physical activity in cardiovascular disease rehabilitation: cluster randomized controlled trial, a pilot study. *JMIR Rehabilitation and Assistive Technologies*, 8(2):e20299, 2021.

[44] J Arthur Harris and Francis G Benedict. A biometric study of human basal metabolism. *Proceedings of the National Academy of Sciences*, 4(12):370–373, 1918.

[45] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[46] André Henriksen, Martin Haugen Mikalsen, Ashenafi Zebene Woldaregay, Miroslav Muzny, Gunnar Hartvigsen, Laila Arnesdatter Hopstock, and Sameline Grimsgaard. Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables. *Journal of medical Internet research*, 20(3):e110, 2018.

[47] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 603–618, 2017.

[48] Yu-Jin Hong, Ig-Jae Kim, Sang Chul Ahn, and Hyoung-Gon Kim. Activity recognition using wearable sensors for elder care. In *2008 second international conference on future generation communication and networking*, volume 2, pages 302–305. IEEE, 2008.

[49] Yan Huang, David Evans, Jonathan Katz, and Lior Malka. Faster secure two-party computation using garbled circuits. In *USENIX security symposium*, volume 201, pages 331–335, 2011.

[50] Zan Huang, Wingyan Chung, and Hsinchun Chen. A graph model for e-commerce recommender systems. *Journal of the American Society for information science and technology*, 55(3):259–274, 2004.

[51] Hyunwoo Hwangbo, Yang Sok Kim, and Kyung Jin Cha. Recommendation system development for fashion retail e-commerce. *Electronic Commerce Research and Applications*, 28:94–101, 2018.

[52] Sana Imtiaz, Muhammad Arsalan, Vladimir Vlassov, and Ramin Sadre. Synthetic and private smart health care data generation using gans. In *2021 International Conference on Computer Communications and Networks (ICCCN)*, pages 1–7. IEEE, 2021.

[53] Tanzir Ul Islam, Reza Ghasemi, and Noman Mohammed. Privacy-preserving federated learning model for healthcare data. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0281–0287. IEEE, 2022.

[54] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *USENIX Security Symposium*, 2019.

[55] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[56] Kanitthika Kaewkannate and Soochan Kim. A comparison of wearable fitness devices. *BMC public health*, 16:1–16, 2016.

[57] Mahdokht Kalantari. Consumers' adoption of wearable technologies: literature review, synthesis, and future research agenda. *International Journal of Technology Marketing*, 12(3):274–307, 2017.

[58] Murat Kantarcıoglu, Jaideep Vaidya, and C Clifton. Privacy preserving naive bayes classifier for horizontally partitioned data. In *IEEE ICDM workshop on privacy preserving data mining*, pages 3–9, 2003.

[59] Andrei Kazlouski, Thomas Marchioro, Harry Manifavas, and Evangelos Markatos. Do you know who is talking to your wearable smartband? *Integrated Citizen Centered Digital Health and Social Care*, page 142, 2020.

[60] Andrei Kazlouski, Thomas Marchioro, Harry Manifavas, and Evangelos P Markatos. I still see you! inferring fitness data from encrypted traffic of wearables. In *HEALTHINF*, pages 369–376, 2021.

[61] Andrei Kazlouski, Thomas Marchioro, and Evangelos Markatos. I just wanted to track my steps! blocking unwanted traffic of fitbit devices. In *Proceedings of the 12th International Conference on the Internet of Things*, pages 96–103, 2022.

[62] Andrei Kazlouski, Thomas Marchioro, and Evangelos P. Markatos. What your fitbit says about you: De-anonymizing users in lifelogging datasets. In *International Conference on Security and Cryptography*, 2022.

[63] Sarah Kozey Keadle, Leah Meuter, Suzanne Phelan, and Siobhan M Phillips. Charity-based incentives motivate young adult cancer survivors to increase physical activity: a pilot randomized clinical trial. *Journal of Behavioral Medicine*, 44:682–693, 2021.

[64] Daniel Kelly, Kevin Curran, and Brian Caulfield. Automatic prediction of health status using smartphone-derived behavior profiles. *IEEE journal of biomedical and health informatics*, 21(6):1750–1760, 2017.

[65] Kathrine Kelly-Schuette, Tamer Shaker, Joseph Carroll, Alan T Davis, G Paul Wright, and Mathew Chung. A prospective observational study comparing effects of call schedules on surgical resident sleep and physical activity using the fitbit. *Journal of Graduate Medical Education*, 13(1):113–118, 2021.

[66] Razaullah Khan, Xiaofeng Tao, Adeel Anjum, Tehsin Kanwal, Saif Ur Rehman Malik, Abid Khan, Waheed Ur Rehman, and Carsten Maple. θ-sensitive k-anonymity: An anonymization model for iot based electronic health records. *Electronics*, 9(5):716, 2020.

[67] Tae Kyun Kim. T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540–546, 2015.

[68] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[69] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[70] Sang-Ho Lee, Yeongmi Ha, Mira Jung, Seungkyoung Yang, and Won-Seok Kang. The effects of a mobile wellness intervention with fitbit use and goal setting for workers. *Telemedicine and e-Health*, 25(11):1115–1122, 2019.

[71] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings 10*, pages 4–15. Springer, 1998.

[72] Hongtao Li, Feng Guo, Wenyin Zhang, Jie Wang, and Jinsheng Xing. (a, k)-anonymous scheme for privacy-preserving data collection in iot-based healthcare services systems. *Journal of Medical Systems*, 42:1–9, 2018.

[73] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

[74] Tong Li, Jin Li, Zheli Liu, Ping Li, and Chunfu Jia. Differentially private naive bayes learning over multiple data sources. *Information Sciences*, 444:89–104, 2018.

[75] Xueping Liang, Sachin Shetty, Juan Zhao, Daniel Bowden, Danyi Li, and Jihong Liu. Towards decentralized accountability and self-sovereignty in healthcare systems. In *Information and Communications Security: 19th International Conference, ICICS 2017, Beijing, China, December 6-8, 2017, Proceedings 19*, pages 387–398. Springer, 2018.

[76] Fang Liu and Tong Li. A clustering k-anonymity privacy-preserving method for wearable iot devices. *Security and Communication Networks*, 2018:1–8, 2018.

[77] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. Protecting sensory data against sensitive inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems*, pages 1–6, 2018.

[78] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*, pages 49–58, 2019.

[79] Thomas Marchioro, Lodovico Giaretta, Evangelos Markatos, and Šarunas Girdzijauskas. Federated naive bayes under differential privacy. In *19th International Conference on Security and Cryptography (SECRYPT), JUL 11-13, 2022, Lisbon, Portugal*, pages 170–180. Scitepress, 2022.

[80] Thomas Marchioro, Andrei Kazlouski, and Evangelos Markatos. How to publish wearables' data: Practical guidelines to protect user privacy. *Stud. Health Technol. Inform*, 294:949–950, 2022.

[81] Thomas Marchioro, Andrei Kazlouski, and Evangelos P Markatos. User identification from time series of fitness data. In *SECRYPT*, pages 806–811, 2021.

[82] Thomas Marchioro, Andrei Kazlouski, and Evangelos P Markatos. Practical crowdsourcing of wearable iot data with local differential privacy. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, pages 275–287, 2023.

[83] Thomas Marchioro, Nicola Laurenti, and Deniz Gündüz. Adversarial networks for secure wireless communications. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8748–8752. IEEE, 2020.

[84] Grace McKeon, Zachary Steel, Ruth Wells, Jill Newby, Dusan Hadzi-Pavlovic, Davy Vancampfort, and Simon Rosenbaum. A mental health–informed physical activity intervention for first responders and their partners delivered using facebook: mixed methods pilot study. *JMIR Formative Research*, 5(4):e23432, 2021.

[85] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[86] Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30, 2009.

[87] Milad Asgari Mehrabadi, Iman Azimi, Fatemeh Sarhaddi, Anna Axelin, Hannakaisa Niela-Vilén, Saana Myllyntausta, Sari Stenholm, Nikil Dutt, Pasi Liljeberg, Amir M Rahmani, et al. Sleep tracking of a commercially available smart ring and smartwatch against medical-grade actigraphy in everyday settings: instrument validation study. *JMIR mHealth and uHealth*, 8(11):e20465, 2020.

[88] Subhas Chandra Mukhopadhyay. Wearable sensors for human activity monitoring: A review. *IEEE sensors journal*, 15(3):1321–1330, 2014.

[89] Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*, 2006.

[90] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[91] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *2009 30th IEEE symposium on security and privacy*, pages 173–187. IEEE, 2009.

[92] Aravind Natarajan. Heart rate variability during mindful breathing meditation. *Frontiers in Physiology*, 13:2792, 2023.

[93] Marlis Ontivero-Ortega, Agustin Lage-Castellanos, Giancarlo Valente, Rainer Goebel, and Mitchell Valdes-Sosa. Fast gaussian naïve bayes for searchlight classification analysis. *Neuroimage*, 163:471–479, 2017.

[94] OpenHumans. Open humans fitbit connection. `https://www.openhumans.org/activity/fitbit-connection`, February 2016.

[95] Ruairi O'Driscoll, Jake Turicchi, Mark Hopkins, Graham W Horgan, Graham Finlayson, and James R Stubbs. Improving energy expenditure estimates from wearable devices: A machine learning approach. *Journal of Sports Sciences*, 38(13):1496–1505, 2020.

[96] B Padmaja, VV Rama Prasad, KVN Sunitha, N Chandra Sekhar Reddy, and CH Anil. Detectstress: A novel stress detection system based on smartphone and wireless physical activity tracker. In *First International Conference on Artificial Intelligence and Cognitive Computing: AICC 2018*, pages 67–80. Springer, 2019.

[97] Abhinav Parate, Meng-Chieh Chiu, Chaniel Chadowitz, Deepak Ganesan, and Evangelos Kalogerakis. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, pages 149–161, 2014.

[98] Juha Parkka, Miikka Ermes, Panu Korpipaa, Jani Mantyjarvi, Johannes Peltola, and Ilkka Korhonen. Activity classification using realistic data from wearable sensors. *IEEE Transactions on information technology in biomedicine*, 10(1):119–128, 2006.

[99] Shyamal Patel, Hyung Park, Paolo Bonato, Leighton Chan, and Mary Rodgers. A review of wearable sensors and systems with application in rehabilitation. *Journal of neuroengineering and rehabilitation*, 9(1):1–17, 2012.

[100] Ignacio Perez-Pozuelo, Bing Zhai, Joao Palotti, Raghvendra Mall, Michaël Aupetit, Juan M Garcia-Gomez, Shahrad Taheri, Yu Guan, and Luis Fernandez-Luque. The future of sleep health: a data-driven revolution in sleep science and medicine. *NPJ digital medicine*, 3(1):42, 2020.

[101] Siobhan Phillips, Payton Solk, Whitney Welch, Lisa Auster-Gussman, Marilyn Lu, Erin Cullather, Emily Torre, Madelyn Whitaker, Emily Izenman, Jennifer La, et al. A technology-based physical activity intervention for patients with

metastatic breast cancer (fit2thrivemb): protocol for a randomized controlled trial. *JMIR Research Protocols*, 10(4):e24254, 2021.

[102] Bernardine M Pinto, Madison Kindred, Regina Franco, Virginia Simmons, and James Hardin. A 'novel' multi-component approach to promote physical activity among older cancer survivors: a pilot randomized controlled trial. *Acta Oncologica*, 60(8):968–975, 2021.

[103] Kedar Potdar, Taher S Pardawala, and Chinmay D Pai. A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, 175(4):7–9, 2017.

[104] KR Pradeep and NC Naveen. Lung cancer survivability prediction based on performance using classification techniques of support vector machines, c4. 5 and naive bayes algorithms for healthcare analytics. *Procedia computer science*, 132:412–420, 2018.

[105] Keerthana Rajendran, Manoj Jayabalan, and Muhammad Ehsan Rana. A study on k-anonymity, l-diversity, and t-closeness techniques. *IJCSNS*, 17(12):172, 2017.

[106] Nisarg Raval, Ashwin Machanavajjhala, and Jerry Pan. Olympus: Sensor privacy through utility aware obfuscation. *Proc. Priv. Enhancing Technol.*, 2019(1):5–25, 2019.

[107] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[108] Mickael Ringeval, Gerit Wagner, James Denford, Guy Paré, and Spyros Kitsiou. Fitbit-based interventions for healthy lifestyle outcomes: systematic review and meta-analysis. *Journal of medical Internet research*, 22(10):e23954, 2020.

[109] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[110] Paul R Rosenbaum, P Rosenbaum, and Briskman. *Design of observational studies*, volume 10. Springer, 2010.

[111] Farida Sabry, Tamer Eltaras, Wadha Labda, Khawla Alzoubi, and Qutaibah Malluhi. Machine learning for healthcare wearable devices: the big picture. *Journal of Healthcare Engineering*, 2022, 2022.

[112] Munshi Saifuzzaman, Tajkia Nuri Ananna, Mohammad Jabed Morshed Chowdhury, Md Sadek Ferdous, and Farida Chowdhury. A systematic literature review on wearable health data publishing under differential privacy. *International Journal of Information Security*, 21(4):847–872, 2022.

[113] Margarita Santiago-Torres, Isobel Contento, Pamela Koch, Wei-Yann Tsai, Adam M Brickman, Ann Ogden Gaffney, Cynthia A Thomson, Tracy E Crane, Naxielly Dominguez, Jhack Sepulveda, et al. ¡Mi Vida Saludable! A randomized, controlled, 2× 2 factorial trial of a diet and physical activity intervention among Latina breast cancer survivors: Study design and methods. *Contemporary Clinical Trials*, 110:106524, 2021.

[114] Aarti Sathyanarayana, Shafiq Joty, Luis Fernandez-Luque, Ferda Ofli, Jaideep Srivastava, Ahmed Elmagarmid, Teresa Arora, and Shahrad Taheri. Sleep quality prediction from wearable data using deep learning. *JMIR mHealth and uHealth*, 4(4):e125, 2016.

[115] Yash Shah, Jocelyn Dunn, Erich Huebner, and Steven Landry. Wearables data integration: Data-driven modeling to adjust for differences in jawbone and fitbit estimations of steps, calories, and resting heart-rate. *Computers in Industry*, 86:72–81, 2017.

[116] Atul Sharma, Mihaela Badea, Swapnil Tiwari, and Jean Louis Marty. Wearable biosensors: an alternative and practical approach in healthcare and disease monitoring. *Molecules*, 26(3):748, 2021.

[117] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-iid data. *arXiv preprint arXiv:1910.07796*, 2019.

[118] Viktoriia Shubina, Sylvia Holcer, Michael Gould, and Elena Simona Lohan. Survey of decentralized solutions with mobile devices for user location tracking, proximity detection, and contact tracing in the covid-19 era. *Data*, 5(4):87, 2020.

[119] Ralf C Staudemeyer and Eric Rothstein Morris. Understanding LSTM–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*, 2019.

[120] Ashleigh Sushames, Andrew Edwards, Fintan Thompson, Robyn McDermott, and Klaus Gebel. Validity and reliability of fitbit flex for step count, moderate to vigorous physical activity and activity energy expenditure. *PloS one*, 11(9):e0161224, 2016.

[121] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

[122] Vajira Thambawita, Steven Hicks, Hanna Borgli, Svein A Pettersen, Dag Johansen, Håvard Johansen, Tomas Kupka, Håkon K Stensland, Debesh Jha, Tor-Morten Grønli, and et al. Pmdata: A sports logging dataset, Feb 2020.

[123] Ilaria Torre, Odnan Ref Sanchez, Frosina Koceva, and Giovanni Adorni. Supporting users to take informed decisions on privacy settings of personal devices. *Personal and Ubiquitous Computing*, 22(2):345–364, 2018.

[124] Ardhendu Tripathy, Ye Wang, and Prakash Ishwar. Privacy-preserving adversarial networks. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 495–505. IEEE, 2019.

[125] Dane Troyer, Justin Henry, Hoda Maleki, Gokila Dorai, Bethany Sumner, Gagan Agrawal, and Jon Ingram. Privacy-preserving framework to facilitate shared data access for wearable devices. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2583–2592. IEEE, 2021.

[126] Jaideep Vaidya and Chris Clifton. Privacy preserving naive bayes classifier for vertically partitioned data. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 522–526. SIAM, 2004.

[127] Jaideep Vaidya, Basit Shafiq, Anirban Basu, and Yuan Hong. Differentially private naive bayes classification. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 571–576. IEEE, 2013.

[128] Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. Collecting and analyzing multidimensional data with local differential privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 638–649. IEEE, 2019.

[129] Bruce J West and Michael Shlesinger. The noise in natural phenomena. *American Scientist*, 78(1):40–45, 1990.

[130] Chugui Xu, Ju Ren, Deyu Zhang, Yaoxue Zhang, Zhan Qin, and Kui Ren. Ganobfuscator: Mitigating information leakage under gan via differential privacy. *IEEE Transactions on Information Forensics and Security*, 14(9):2358–2371, 2019.

[131] Mengmeng Yang, Lingjuan Lyu, Jun Zhao, Tianqing Zhu, and Kwok-Yan Lam. Local differential privacy and its applications: A comprehensive survey. *arXiv preprint arXiv:2008.03686*, 2020.

[132] Sofia Yfantidou, Christina Karagianni, Stefanos Efstathiou, Athena Vakali, Joao Palotti, Dimitrios Panteleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Šarūnas Girdzijauskas. Lifesnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Scientific Data*, 9(1):663, 2022.

[133] Sofia Yfantidou, Pavlos Sermpezis, and Athena Vakali. Self-tracking technology for mhealth: A systematic review and the past self framework. *arXiv preprint arXiv:2104.11483*, 2021.

[134] Xun Yi and Yanchun Zhang. Privacy-preserving naive bayes classification on distributed data via semi-trusted mixers. *Information systems*, 34(3):371–380, 2009.

[135] Xuefei Yin, Yanming Zhu, and Jiankun Hu. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)*, 54(6):1–36, 2021.

[136] Amir Zadeh, David Taylor, Margaret Bertsos, Timothy Tillman, Nasim Nosoudi, and Scott Bruce. Predicting sports injuries with wearable technology and data analysis. *Information Systems Frontiers*, 23:1023–1037, 2021.

[137] Haibin Zhang, Bo Wen, Jiajia Liu, and Yingming Zeng. The prediction and error correction of physiological sign during exercise using bayesian combined predictor and naive bayesian classifier. *IEEE Systems Journal*, 13(4):4410–4420, 2019.

[138] Harry Zhang. The optimality of naive bayes. *Aa*, 1(2):3, 2004.

[139] Zhenjiang Zhang, Bowen Han, Han-Chieh Chao, Feng Sun, Lorna Uden, and Di Tang. A new weight and sensitivity based variable maximum distance to average vector algorithm for wearable sensor data privacy protection. *IEEE Access*, 7:104045–104056, 2019.

[140] Yang Zhao, Jun Zhao, Mengmeng Yang, Teng Wang, Ning Wang, Lingjuan Lyu, Dusit Niyato, and Kwok-Yan Lam. Local differential privacy-based federated learning for internet of things. *IEEE Internet of Things Journal*, 8(11):8836–8853, 2020.