Practical Crowdsourcing of Wearable IoT Data with Local Differential Privacy

Thomas Marchioro*
marchiorot@ics.forth.gr
Foundation for Research and Technology Hellas
Heraklion, Crete, Greece

Andrei Kazlouski*
andrei@ics.forth.gr
Foundation for Research and Technology Hellas
Heraklion, Crete, Greece

Evangelos Markatos
markatos@ics.forth.gr
Foundation for Research and Technology Hellas
Heraklion, Crete, Greece

ABSTRACT

In this work, we present and evaluate a crowdsourcing platform to collect wearable IoT data with local differential privacy (LDP). LDP protects privacy by perturbing data with noise, which may hinder their utility in some cases. For this reason, most researchers are wary of adopting it in their studies. To address these concerns, we consider the impact of different privacy budget values on the real wearable IoT data (steps, calories, distance, etc.) from $N = 71$ Fitbit users. Our goal is to demonstrate that, even if the collected information is protected with LDP, it is possible for data analysts to extract statistically significant insights on the studied population. To this end, we evaluate the error for various metrics of interest, such as sample average and empirical distribution. Furthermore, we verify that, in most cases, statistical tests produce the same results regardless of whether LDP has been applied or not. Our findings suggest that LDP with a privacy budget between 4 and 8 maintains an acceptable error of $\leq 3\%$ and over 90% agreement on t-tests. Finally, we show that such values of privacy budget, albeit providing loose theoretical guarantees, can effectively defend against re-identification attacks on wearable IoT data.

CCS CONCEPTS

• **Security and privacy**; • **Human-centered computing** → *Ubiquitous and mobile computing*; • **Applied computing** → *Life and medical sciences*;

ACM Reference Format:
Thomas Marchioro, Andrei Kazlouski, and Evangelos Markatos. 2023. Practical Crowdsourcing of Wearable IoT Data with Local Differential Privacy. In *International Conference on Internet-of-Things Design and Implementation (IoTDI '23), May 09–12, 2023, San Antonio, TX, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3576842.3582367

1 INTRODUCTION

Consumer-level fitness trackers provide not only a non-invasive tool for physical self-assessment [6] but also a valuable resource in medical research [13, 30]. 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, researchers can select candidates who already own a suitable wearable tracker. This way, analysts do not need to buy new devices and to meet participants in person. The spread of COVID-19 has already established crowdsourcing of data produced by fitness trackers as a valuable tool to detect the virus and combat the pandemic [28]. However, the systematic collection of such data, known in literature as "wearable IoT data", has raised several concerns in later years, due to their sensitive nature. Readily available crowdsourcing platforms tend to put less emphasis on privacy, while focusing more on simplifying the data collection process. In the case of wearable IoT data, neglecting the privacy aspect may lead to re-identification of individuals who want to remain anonymous. This is particularly problematic in health studies, which typically target people with sensitive conditions.

In this paper, we propose and evaluate a simple design for a privacy-preserving crowdsourcing platform based on local differential privacy (LDP) [34]. LDP is a well-established technique to protect sensitive data through random perturbations, making datapoints statistically indistinguishable. In our crowdsourcing setting, participants who connect to the platform can use LDP to produce anonymous reports, thereby safeguarding themselves against re-identification threats. Unfortunately, the protection provided by LDP comes with a price. High levels of noise introduce error in the data, possibly hindering their utility. In LDP, the privacy-utility tradeoff is regulated by a parameter called "privacy budget". Smaller values of privacy budget offer stronger privacy guarantees, but also require to inject more noise in the data. Due to the uncertainty of data quality, health researchers tend to be wary of using LDP in their studies.

Motivated by such concerns, we assess the usability of wearable IoT data collected under LDP. We evaluate how accurately an analyst can estimate various metrics of interest, namely average, inverse cumulative distribution, and statistical significance (in terms of p-value). The latter aspect is particularly relevant, since the vast majority of the medical studies are based on randomized controlled trials, in which determining the statistical significance of the results is a primary objective [17]. This is typically done via statistical tests, such as Student's test, which compares the difference between two groups and determines the likelihood of observing this difference by chance. Therefore, we specifically focus on preserving the results of these tests after applying LDP to the raw data.

Furthermore, by incorporating a third-party server into our platform design, we achieve a more favorable privacy-utility tradeoff

*Both authors contributed equally to this research.

IoTDI '23, May 09–12, 2023, San Antonio, TX, USA

ACM ISBN 979-8-4007-0037-8/23/05.
https://doi.org/10.1145/3576842.3582367

for LDP. This server, enables participants to submit multiple "independent" reports without consuming additional privacy budget. Also, it brings an additional layer of anonymity between the participants and the analyst who is conducting the study. In summary, the contributions of this paper are as follows:

- We present a simple crowdsourcing platform design to collect anonymous reports of wearable IoT data. This design enables analysts to collect multiple reports from the same group of participants without consuming additional privacy budget.
- We show that LDP can be used to mitigate re-identification attacks, both theoretically and through empirical evaluation.
- We estimate the impact of noise on aggregated metrics by testing known LDP mechanisms on a real dataset of $N = 71$ Fitbit users.
- We demonstrate that properly calibrated LDP preserves statistical significance of the obtained results, which is assessed by examining the p-value obtained from Student's tests. We measure the proportion of samples that produces equivalent levels of significance before and after applying LDP. Throughout the paper, we refer to this proportion as *agreement rate*.

Our results suggest that for a sufficient number of participants (e.g., $N = 30$) LDP with loose theoretical guarantees provides a reasonable privacy-utility tradeoff in practice. For instance, it is possible to estimate average reported values (steps, calories, etc.) with less than 3% error and over 90% agreement rate on t-tests, utilizing LDP with the Laplace mechanism and privacy budget $\varepsilon = 8$. The same amount of noise brings the accuracy of a linking attack close to a random guessing strategy. All the experiments reported in the paper are reproducible, and the code is publicly available at the following URL: https://github.com/thomasmarchioro3/CrowdsourcingWearableLDP.

The rest of the paper is structured as follows. In section 2, we introduce the notation used in the paper, along with fundamental concepts such as local differential privacy and independent t-tests. In section 3 we provide an overview of the related work. In section 4, we present a simple crowdsourcing setting that is used to guarantee the anonymity of submissions and to allow analysts to collect multiple reports with a fixed privacy budget. In section 5 we introduce quality and privacy metrics that are used to evaluate the tradeoff offered by local differential privacy. Furthermore, we derive estimators that are used to compute metrics of interest from the collected anonymous reports. We present the results of our experiments in section 6. Finally, we discuss the overall usability of local differential privacy for crowdsourcing wearable data in section 7 and draw the conclusions in section 8.

2 BACKGROUND AND NOTATION

Throughout this paper, we write random variables (r.v.s) in upper case and their observations with the relative lower case letter. We treat the original records $X_1, \dots, X_N$ and the corresponding anonymous reports[1] $Y_1, \dots, Y_N$ as r.v.s with unknown underlying distribution. The reason for using a r.v. model is that, if the LDP mechanism is fixed, the conditional distribution of Y_i given X_i is known. We leverage this distribution to derive estimators for the metrics of interest and an optimal attack strategy for participant re-identification. The observations $y_1, \dots, y_N$ instead are the actual values of the data points collected by the analyst, which are used as input for the estimators to compute the metrics of interest.

[1]We call the randomized version of the fitness records with LDP "anonymous reports" or "anonymous records" interchangeably.

Table 1: Notation used in the paper.

ε	Privacy budget
N	Number of users/participants
m	Number of features in a record
x_i	Original record by user i
y_i	Anonymized record by user i
X_i	Random variable for an original record
Y_i	Random variable for an anonymized record
$\bar{\cdot}$	Sample average
$Q(\cdot)$	Inverse cumulative distribution function (ICDF)
$\Pr[\cdot]$	Probability of an event
$p(\cdot)$	Probability density function (PDF)
$\mathbb{E}[\cdot]$	Expectation
$\chi(\cdot)$	Indicator function
p	p-value

The notation used in this paper is summarized in table 1. It is worth mentioning that the indicator function $\chi(\cdot)$ is used to determine whether to include terms in an equation based on a certain condition. If the condition is satisfied, the indicator function returns a value of 1, implying that the corresponding term should be included. Otherwise, the function returns 0.

In the rest of the section we describe the main concepts used in this paper.

2.1 Local differential privacy

Differential privacy [10] is a mathematical definition of privacy for randomized algorithms. An algorithm that satisfies differential privacy aims to protect individual datapoints in a dataset by perturbing its output with properly calibrated noise. In the traditional differential privacy model, the whole dataset is entrusted to a "curator", who runs some aggregation queries on the data and adds noise to the output. However, if the information in the dataset belongs to various individuals, this centralized approach requires them to disclose their data to the curator. This is not an ideal solution if the data curator is not trusted. In local differential privacy (LDP) [34], on the other hand, datapoints are anonymized at the owner's side, removing the need for a trusted authority. An LDP mechanism $\mathcal{A}$ randomizes an individual datapoint x, outputting an anonymous report Y that should be statistically indistinguishable from any other report. Formally, for any pair of datapoints x, x', the output of the mechanism should satisfy

$$\Pr[\mathcal{A}(x) \in O] \leq e^{\varepsilon} \Pr[\mathcal{A}(x') \in O], \ \forall O \subseteq \text{Range}(\mathcal{A}). \quad (1)$$

The parameter ε is called *privacy budget* and regulates the tradeoff between utility and privacy offered by the mechanism. In this paper, we use LDP to mitigate the effectiveness of re-identification attacks, as explained in section 5.

We utilize two well-known LDP mechanisms: Laplace and Piecewise. Both these mechanisms are defined for scalar inputs and can be extended to multidimensional inputs by applying them separately to each component. In the latter case, the privacy budget must be distributed between the components, according to the sequential composability property of differential privacy [9]. This property states that to enforce ε-differential privacy on an array of m components, one must apply the Laplace or Piecewise mechanism to each component with budget ε/m.

Laplace mechanism. The Laplace mechanism [9] is applied to a scalar value $x \in [x_{\min}, x_{\max}]$ as follows:

$$Y = x + \Xi,\ \Xi \sim \text{Lap}(0, \Delta/\varepsilon). \tag{2}$$

The resulting randomized value is distributed according to $Y \sim \text{Lap}(x, \Delta/\varepsilon)$, and its standard deviation is proportional to the *sensitivity*

$$\Delta = x_{\max} - x_{\min} \tag{3}$$

and inversely proportional to the privacy budget ε. The output of the Laplace mechanism can take any values in $(-\infty, +\infty)$. However, values that are farther from the input range $[x_{\min}, x_{\max}]$ are reached with lower probability.

Piecewise mechanism. The Piecewise mechanism was originally introduced by [33] and improved by [37]. The core idea of the mechanism is to randomize an input $x \in [-1, 1]$ to a limited range $[-A, A]$ 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 PDF is described by the following equation:

$$p(y|x) = \frac{\tau(e^\varepsilon - 1)}{2(\tau + e^\varepsilon)^2} \begin{cases} e^\varepsilon, & \text{if } y \in (L(x), R(x)) \\ 1, & \text{if } y \in [-A, L(x)] \cup [R(x), A] \end{cases} \tag{4}$$

where

$$\begin{aligned} A &= \frac{(e^\varepsilon + \tau)(\tau + 1)}{\tau(e^\varepsilon - 1)},\ L(x) = \frac{(e^\varepsilon + \tau)(x\tau - 1)}{\tau(e^\varepsilon - 1)}, \\ R(x) &= \frac{(e^\varepsilon + \tau)(x\tau + 1)}{\tau(e^\varepsilon - 1)}. \end{aligned} \tag{5}$$

Normally, τ is suitably chosen depending on the values of x and ε. However, in our experiments we adopt the sub-optimal solution $\tau = e^{\varepsilon/3}$ [37]. 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. 4-5, and then rescaled back. Since the Piecewise mechanism outputs values in $[-A, A]$, the rescaled output falls in the range $[x_{\min} + x_{\max}\frac{1-A}{2}, x_{\min} + x_{\max}\frac{1+A}{2}]$.

2.2 Independent t-test

An independent Student's t-test [22] is a statistical hypothesis test that compares the averages of samples collected by two distinct populations. The outcome of a t-test allows to determine if the difference between the average values is statistically significant based on the resulting p-value. Given two groups A and B of equal size $N/2$, the test requires to compute a t statistic as

$$t = \frac{\bar{x}_A - \bar{x}_B}{s/\sqrt{N/2}} \tag{6}$$

where $\bar{x}_A$ and $\bar{x}_B$ are the sample average of the respective populations, and s is the overall sample standard deviation. The value of t and the size of the populations are used to compute the p-value, which essentially represents the likelihood for the two populations to follow the same distribution. When the p-value is below a suitably-chosen threshold α, it means that the difference between the two populations is statistically significant. A typical threshold is $\alpha = 0.05$. In the rest of the paper, for the sake of clarity, we say that a t-test is "passed" if $p < \alpha$, i.e. if the two populations are distinct, and "failed" otherwise.

3 RELATED WORK

Previous studies conducted experiments with Fitbit devices, analyzed LDP and other ways to modify fitness samples in the context of IoT, and researched the possible privacy risks associated with wearable data.

Randomized controlled trials and crowdsourcing of IoT data. Increasing popularity of IoT fitness trackers prompted a significant number of works adopting them. Hence, Fitbit opened its own research library, counting over 1000 articles that utilize wearables [11]. Lee et al. studied the benefits of Fitbit-aided wellness interventions on workers [23]. Wearables were also used to monitor physical activity of patients with hypertension in [3]. Kim et al. conducted an observation work [20], regarding the use of physical activity interventions to prevent metabolic syndrome. By utilizing Fitbit devices, a number of studies collected lifelogging dataset where users wore trackers for fixed periods of time [14, 27, 32]. Finally, as mentioned in the introduction, crowdsourcing of wearable IoT data without any privacy guarantees is already available via online platforms such as Amazon Mechanical Turk [12] and Open Humans [27].

Sanitization of IoT data. DP [9, 10] has long been a standard solution to release all kinds of sensitive data. However, the original DP setting requires all the data to be entrusted to a single curator, which is not a viable solutions in many applications. Therefore, a number of works have studied LDP in the context of IoT data. A literature review by Saifuzzaman et al. investigated the applicability of DP and LDP for wearable IoT data [31]. In [35] authors proposed relaxed definitions of LDP for IoT data, which, nonetheless, do not provide the same guarantees as differential privacy. Another prominent approach for protecting IoT data that has been gaining recognition recently is adversarial machine learning [4, 24, 25]. Malekzadeh et al. managed to preserve the utility of sensor (accelerometer and gyroscope) data, while removing the author- and demographic-specific traits for a 24-users dataset. Their findings were slightly improved in [4]. Although the obtained results are promising, these machine learning-based approaches rely on data-driven models, which are trained on specific types of sensor records. The generalization capability of such models to other tasks and datasets remains unclear. Imitaz et al. [15] applied generative adversarial networks (GANs) coupled with DP to produce fitness samples based on a real-world

Fitbit dataset. Nevertheless, the authors neither evaluated the resilience generated data against realistic privacy threats or assessed their utility. Other studies investigated the possibility of collecting anonymous step count reports and measured how accurately one can estimate the average [21]. To our knowledge, this work is the first one to use the agreement in statistical tests to measure the utility of sanitized wearable IoT data. While prior research focused on general utility metrics, our contribution aims to determine the usability of these data in health research, in which statistical significance is a primary concern. To this end, we assess if LDP changes the outcome of t-tests (i.e., from "passed" to "failed" and vice versa).

Privacy risks and attacks on wearable IoT data. Previous studies have indicated that a number of significant insights may be inferred from IoT sensor data, including food intake [8], smoking [29], health status [19], and even Covid-19 trends [38]. The possibility of re-identifying an individual based on wearable IoT data has been discussed since the early spread of smart fitness trackers [7]. In prior works of ours, we evaluated the effectiveness of linking attacks on wearable IoT data [18, 26]. These works mainly focus on a scenario where the attacker has access to additional records produced by the target user. Malekzadeh et al. showed that users can be de-anonymized [25] and their physical parameters, such as gender, age, and height, can be inferred [24] based on the sensor data of mobile phones.

Specifics of LDP. Most of the LDP implementations rely on either noise-based mechanisms (e.g., Laplace and Gaussian) or randomized response (e.g., RAPPOR) [1]. Randomized response is preferable for time series of IoT data streams with low granularity and high sampling rate, e.g., heart rate measurements [2]. Alternatively, noise-based mechanisms are more suitable to maintain utility of the individual values. In our work, we focus on the latter, since the metrics of interest are computed on individual datapoints in our crowdsourcing scenario.

4 CROWDSOURCING SETTING

In this section, we describe our design for a privacy-preserving crowdsourcing platform to collect anonymous reports under LDP. We first detail the platform requirements, which should meet the needs of both the data providers (i.e., the participants of the study) and the analyst. Such requirements are:

(1) *Anonymity*: The analyst – and any other entity who has access to the randomized data – must not know which participant has sent an anonymous report.
(2) *Quality*: The analyst must be able to use the anonymous reports to compute aggregated metrics of interest (average, ICDF, p-value of statistical tests) within a certain margin of error.
(3) *Accountability*: The analyst must be able to reward participants when they send an anonymous report. Conversely, a participant who does not submit any data should not be rewarded.

The experimental results discussed in the next sections hold for any crowdsourcing platform design that complies with the above requirements. Our proposed solution involves multiple participants and an analyst communicating using a third-party server as intermediary, as depicted in figure 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[2] 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.
- 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 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

[2]We refer to the analyst as she/her, and to a participant as he/him.

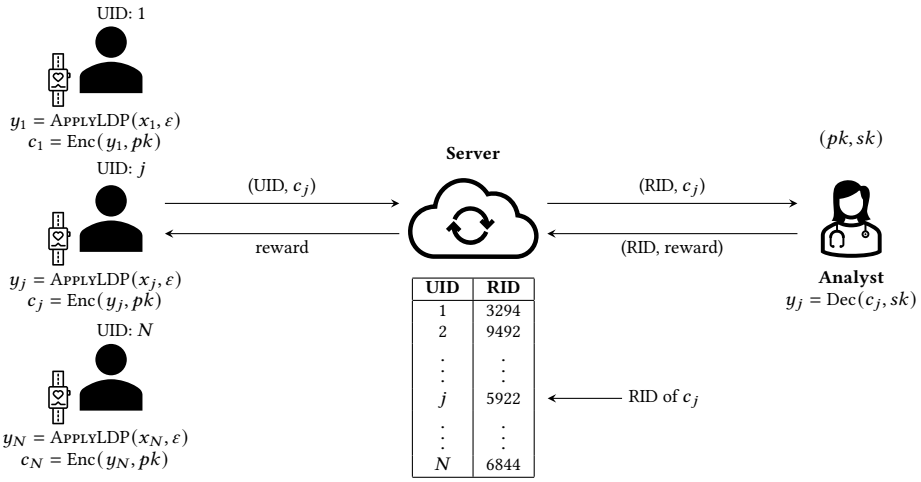

Figure 1: Simple design of a crowdsourcing platform that allows to submit abobymous reports under LDP guarantees. Users submit reports of wearable IoT data once per day. A participant with user identifier (UID) j randomizes his daily report under ε-LDP and encrypts it with a public key pk. Then, he sends the pair (UID, c_j) to a third-party crowdsourcing server that assigns a random report identifier (RID) to c_j. The server forwards the pair (RID, c_j) to the analyst, who decrypts the report with a secret key sk and sends back a reward for the corresponding RID. The server forwards the reward to the correct user. Unless the analyst and the server conspire against a user, neither can compromise his anonymity.

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.

5 METHODOLOGY

In order to satisfy the "quality" requirement (section 4), we need to ensure that the analyst is able to accurately compute a number of metrics of interest. In this section, we derive estimators to compute sample average, inverse cumulative distribution function (ICDF), and p-value of t-tests based on noisy samples $y_1, \ldots, y_N$. The ability to calculate these metrics on a daily basis allows to monitor the progress of participants during rehabilitation or physical activity intervention. The sample average can be used to compare two populations, e.g., to determine if a certain strategy would encourage participants to take more steps. The p-value allows to determine if this comparison is statistically significant. 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., $x = 10,000$.

5.1 Estimators under LDP

Since LDP introduces noise in the reported samples, such modifications should be taken into account when estimating the metrics of interest. In the following paragraphs, we define and motivate the estimators used for sample average, ICDF, and p-value.

Sample average. The sample average of the original samples $x_1, \ldots, x_N$ is simply $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. The r.v. Y_i, representing the ith randomized report, has mean $\mathbb{E}[Y_i] = x_i$ for both the Laplace and Piecewise mechanisms. Therefore, it holds

$$\mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} Y_i] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[Y_i] = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x} \tag{7}$$

meaning that

$$\hat{\theta}(\bar{x}) = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y} \tag{8}$$

is an unbiased estimator for the sample average.

Empirical ICDF. The empirical ICDF of $x_1, \ldots, x_N$ is computed for each $x \in \mathcal{X}$ as

$$Q(x) = \frac{1}{N}\sum_{i=1}^{N} \chi\{x_i > x\}. \tag{9}$$

Under the observations $Y_1 = y_1, \ldots, Y_N = y_N$, the empirical ICDF can be estimated as

$$\mathbb{E}[Q(x)|Y_i = y_i] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} \chi\{X_i > x\}|Y_i = y_i] \tag{10}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\chi\{X_i > x\}|Y_i = y_i] \tag{11}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \Pr[X_i > x|Y_i = y_i]. \tag{12}$$

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

becomes

$$\hat{\theta}(Q(x)) = \frac{1}{N}\sum_{i=1}^{N} \Pr[y_i - \Xi_i > x] \tag{13}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \Pr[\Xi_i < y_i - x] \tag{14}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \begin{cases} \frac{1}{2}e^{\varepsilon\frac{y_i - x}{\Delta}} & \text{if } y_i \le x, \\ 1 - \frac{1}{2}e^{\varepsilon\frac{x - y_i}{\Delta}} & \text{otherwise.} \end{cases} \tag{15}$$

Figure 2 shows that eq. 15 can effectively be used to estimate the empirical ICDF.

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 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

$$\hat{\theta}(t) = \frac{\bar{y}_A - \bar{y}_B}{\hat{s}/\sqrt{N}}, \tag{16}$$

where $\hat{s}$ is the overall sample standard deviation. Based on $\hat{\theta}(t)$, we compute the corresponding p-value $\hat{\theta}(p)$. Applying LDP does not introduce a bias in the mean values $\bar{y}_A$ and $\bar{y}_B$, since they can increase or decrease with equal probability. This implies that also the difference $\bar{y}_A - \bar{y}_B$ On the other hand, the sample variance is systematically 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, since overestimating the p-value is preferable to an underestimation, as we explain in section 5.2.

5.2 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 α. 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.

RMSE and NRMSE. For the sample average $\bar{x}$ and the ICDF $Q(x)$, 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. Let $\hat{\theta}(\bar{x}^{(\ell)})$ be the estimated sample average at day ℓ, and let $\bar{x}^{(\ell)}$ be its true value, calculated without applying LDP noise. The RMSE over L days is computed as

$$\text{RMSE} = \sqrt{\frac{1}{L}\sum_{\ell=1}^{L}\left(\bar{x}^{(\ell)} - \hat{\theta}(\bar{x}^{(\ell)})\right)^2}. \tag{17}$$

The calculation is analogous for the ICDF, replacing $\bar{x}$ with $Q(x)$ and $\hat{\theta}(\bar{x})$ with $\hat{\theta}(Q(x))$. We choose RMSE over mean absolute error (MAE), used in other works [31], since RMSE penalizes sporadic large errors, and in randomized controlled trials consistently low errors are desirable. RMSE can also be normalized to express the error in a percent form as

$$\text{NRMSE} = \frac{\text{RMSE}}{x_{\max} - x_{\min}} \tag{18}$$

for the sample average. The RMSE on the ICDF is already normalized w.r.t. the number of users N, so we also call it "NRMSE" in our results.

Agreement rate. 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 α, i.e., $\hat{\theta}(p) < \alpha \Leftrightarrow p < \alpha$. The *agreement rate* over n trials with p-value threshold α is as

$$\frac{1}{n}\sum_{\nu=1}^{n} \chi\{p^{(\nu)} < \alpha \wedge \hat{\theta}(p^{(\nu)}) < \alpha\} + \chi\{p^{(\nu)} \ge \alpha \wedge \hat{\theta}(p^{(\nu)}) \ge \alpha\}, \tag{19}$$

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 2: the type I error (false positive) occurs when $p < \alpha$ but $\hat{\theta}(p) > \alpha$, while type II error occurs in the opposite scenario. While t-tests can demonstrate the difference between 2 populations (if $p < \alpha$), they cannot disprove such difference (if $p > \alpha$). 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 in section 6, where we show that our p-value estimator consistently achieves high-rate agreement when $p > \alpha$ on the original data.

Table 2: Different types of agreement and errors in t-tests under LDP. A standard threshold value is $\alpha = 0.05$, which implies 95% confidence.

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

5.3 Re-identification attack

The purpose of applying LDP is to protect participants against re-identification. To determine the level of protection offered by an

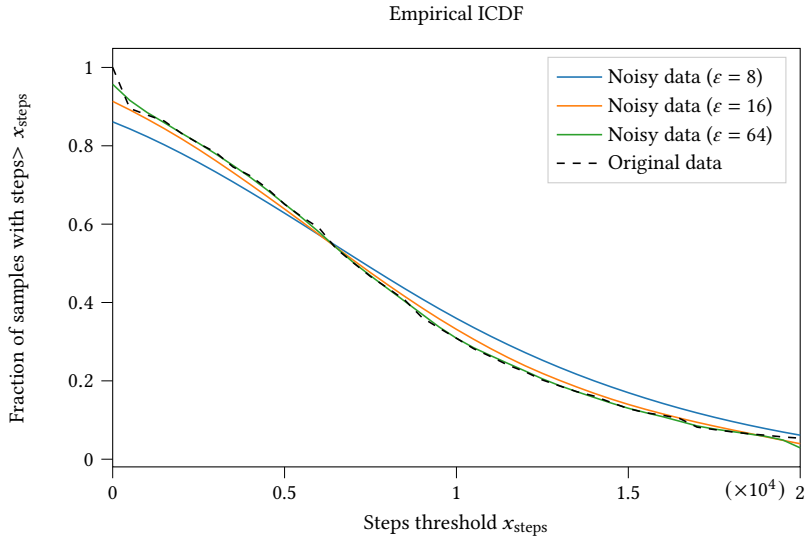

Figure 2: Example of empirical ICDF estimation for different values of ε. Evaluating the ICDF allows to count how many participants achieved a certain step goal.

LDP mechanism, we study its effectiveness against the following threat model, depicted in fig. 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 on the same day. We call this type of re-identification strategy a *linking attack*, as it aims to link the original record to the anonymized record. In this threat model, a naive guessing approach would yield a $1/N$ re-identification probability, meaning that, ideally, LDP-protected records should bring the success rate of the attack close to this value. Contrarily to membership inference attacks, we assume that the adversary knows that the target is present in the dataset of anonymous records. We also assume that the details of the LDP mechanism are known to the adversary.

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 pairs of steps and calories to find individuals with distinct physical characteristic (such as height and weight), since these characteristics are used to estimate calories from steps. Our prior work showed that re-identification based on wearable records can become a realistic threat if the attacker has a reasonable amount of background information on her target [18, 26]. 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.

Linking criterion. 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

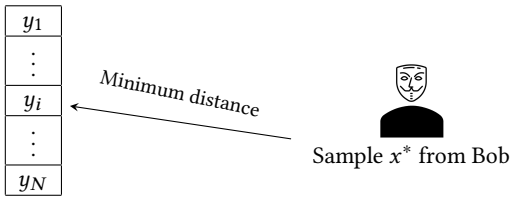

Figure 3: Linking attack considered in our evaluation. The adversary (Eve) aims to re-identify her target (Bob) by leveraging the original sample x^* and comparing it to the anonymized records $y_1, \ldots, y_N$. Eve selects the "closest" record y_i to x^* (according to eq. 20).

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

$$\hat{y} = \arg \min_{y_i, i=1,\ldots,N} \sum_{f=1}^{m} \left(\frac{|y_i[f] - x^*[f]|}{x_{\max}[f] - x_{\min}[f]} \right), \tag{20}$$

where $f = 1, \ldots, m$ 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), as shown in appendix A. For the Piecewise mechanism, the optimal decision is

$$\hat{y} = \arg \max_{y_i, i=1,\ldots,N} \sum_{f=1}^{m} \chi \left\{ y_i[f] \in \left(L(x^*[f]), R(x^*[f]) \right) \right\}, \tag{21}$$

i.e., choosing the report with most features in the high-density regions $(L(x^*[f]), R(x^*[f]))$ for the original data point x^*. However, in most practical cases this is equivalent to the minimum distance criterion. Therefore, in our experiments we adopt the criterion described in eq. 20, since it is faster to evaluate, and thus more suitable for Monte Carlo experiments.

LDP and linking rate. Enforcing LDP on the reports sensibly limits the success probability of a linking attack, as shown by our experimental results in section 6. The level of protection granted by LDP depends on the privacy budget ε, the number of features m, 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 $\Pr[\mathcal{S}|m, N, \varepsilon, \text{Laplace}] = \gamma$,

$$\gamma \leq 1 - e^{-\varepsilon}\left(1 - \left(1 - \left(\frac{1}{2} - \frac{1}{2}e^{-\frac{2\varepsilon}{m}}\right)^m\right)^{N-1}\right); \tag{22}$$

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

$$\gamma' \leq 1 - \frac{e^{\frac{\varepsilon}{3}}}{(e^{\frac{\varepsilon}{3m}} + e^{\frac{\varepsilon}{m}})^m}\left(1 - \left(1 - \frac{1}{(e^{\frac{\varepsilon}{3m}} + e^{\frac{\varepsilon}{m}})^m}\right)^{N-1}\right). \tag{23}$$

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.

5.4 Dataset

We test the effects of LDP on LifeSnaps [36], a real fitness dataset comprising $N = 71$ participants from different European countries. LifeSnaps contains daily records of steps, burned calories, covered distance, and other activity indicators. Such records were collected during two different rounds of 64 days each: the first including 44 participants, and the second covering the remaining 31. For the purpose of our experiments, the samples are aggregated by day and the dates are aligned as if all the participants belonged to the first round. The experiments reported in section 6 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.

6 EXPERIMENTAL RESULTS

In this section we present the results of our experiments conducted on the LifeSnaps dataset. 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 $n = 100$ 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 privacy budget ε;
- We compute the metrics of interest.

Table 3: Minimum and maximum values chosen for each feature. Inputs are clipped in the interval $[x_{\min}, x_{\max}]$ before applying LDP. This allows to compute the sensitivity for the LDP mechanisms.

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

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 fixed privacy budget $\varepsilon = 8$. Due to space limitations, 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 [3, 23]. While estimating the success rate of linking attacks, we consider the combination of steps and calories, since that was proven to be highly identifying [26]. Other features yield similar results.

Input clipping. 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 3.

6.1 Error on aggregated metrics

When computing the error on aggregated metrics, the RMSE is computed taking into account original and anonymized samples across the 64 days in the dataset. Figure 4 shows the RMSE between the true sample mean $\bar{x}$ for steps and the estimate $\hat{\theta}(\bar{x})$, for varying privacy budget and number of participants. The RMSE decreases when ε and/or N are increased. This is expected, since a larger number of records reduce the variance for the sample average estimator. Apparently, the Piecewise mechanism introduces less error than Laplace for a 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 $\varepsilon \geq 8$. That is an acceptable error, about 3% of the overall range $[0, 20000]$. The Piecewise mechanism, on the other hand, reaches the same utility at $\varepsilon = 4$. This must be taken into account when choosing a suitable ε for the anonymous reports.

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 $N \times \text{ICDF}(10000)$. 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. 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., normalized w.r.t. N – decreases with N. This implies that the fraction of participants that met a certain step goal can be estimated with high confidence when the number of participants is large.

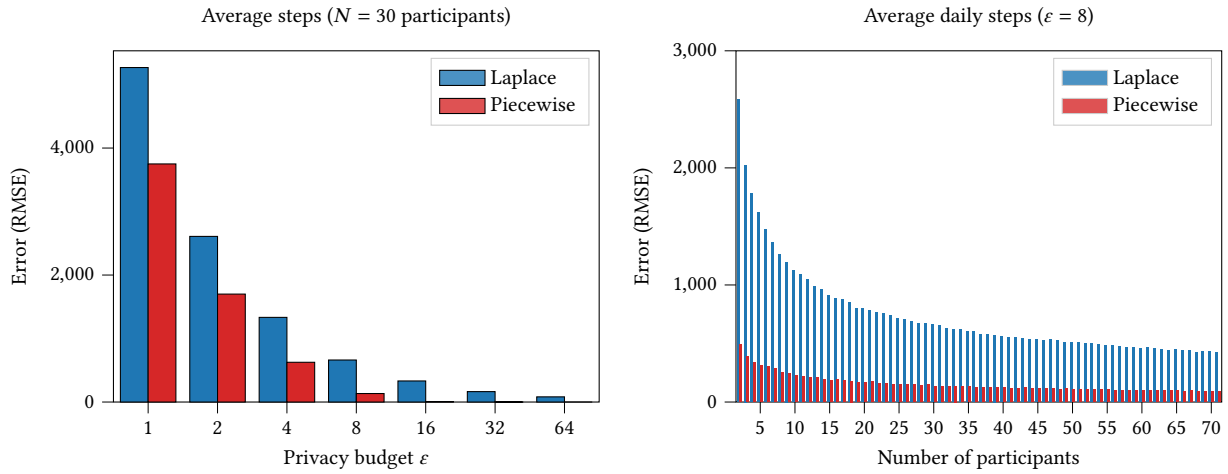

Figure 4: RMSE of average step estimates under LDP for varying number of participants N and privacy budget ε. Unsurprisingly, a larger number of participants provides a more accurate estimation of the average. For a same (N, ε) pair, the Piecewise mechanism introduces less noise.

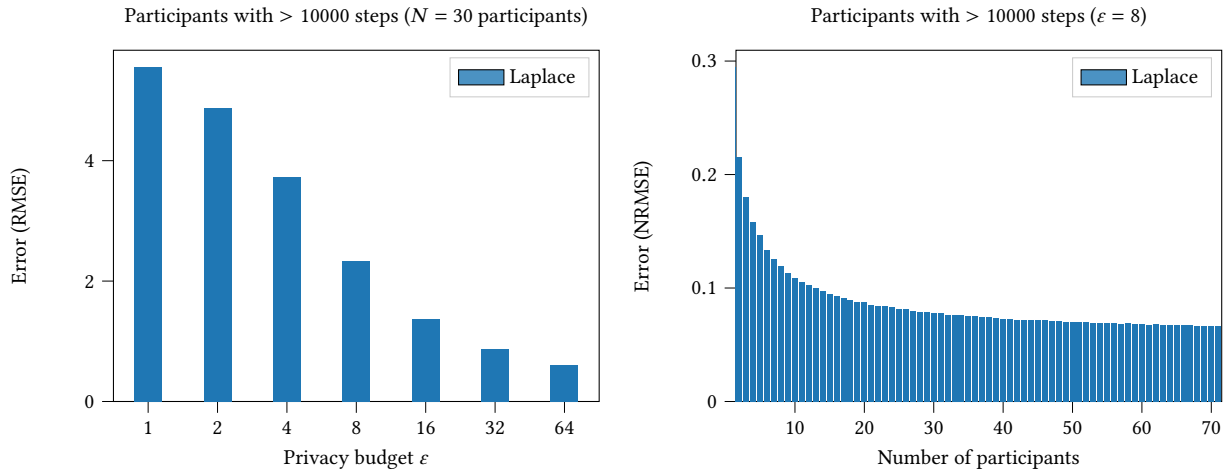

Figure 5: RMSE of count estimate ($N \times$ ICDF) for users taking over 10000 steps per day.

6.2 Agreement on t-tests

The agreement on t-tests is measured by sampling an even number of participants and dividing them into 2 groups of equal size. For example, the left plot in fig. 6 shows the agreement rate for $N = 30$ participants, meaning that 30 users were sampled at random and split in two groups of 15. For this experiment, since two types of error need to be evaluated, we increased the number of iterations of each Monte Carlo experiment to $n = 1000$. The threshold to determine statistical significance was set at $\alpha = 0.05$. In a randomized controlled trial, running a t-test on two groups of participants can either show that their sample averages to be significantly different ($p < 0.05$) or not ($p \geq 0.05$). The agreement rate represents the fraction of t-tests run on anonymous reports to provide the same outcome as the corresponding t-tests run on the original records.

Our choice for the estimator mainly leads to type II errors, meaning that a t-test on noisy reports shows no statistical significance where $p < 0.05$ on the original data. On the other hand, when the original data do not actually show a significant difference between the two groups, the agreement rate is consistently high, regardless of the value of ε. This implies that this estimation approach is robust against type II error as depicted in table 2. As in the sample average case, the Piecewise mechanism provides higher utility than Laplace for a same value of ε. Even more surprisingly, the same values of $\varepsilon = 4$ and $\varepsilon = 8$ appear to work as a good threshold between high- and low-utility results, achieving over 90% agreement rate. Therefore, such values of ε may be the ideal choice for the practical applications.

An interesting finding, shown in the right plot of fig. 6, is that the number of participants seems to have no impact on the results.

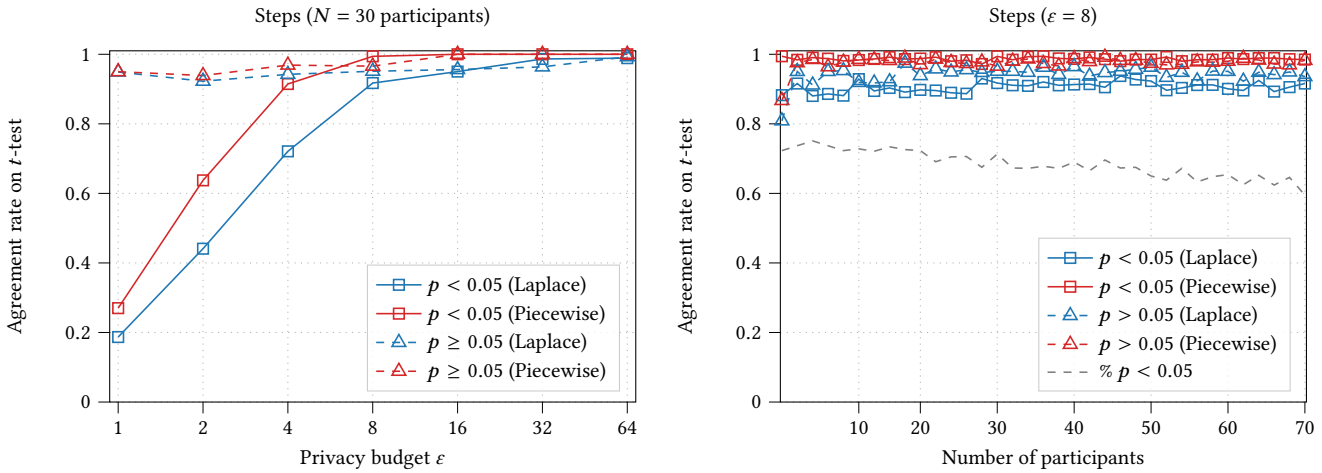

Figure 6: Agreement on t-tests for varying privacy budget and number of participants. The agreement rate is divided between the cases where the original data yield statistically significant results ($p < 0.05$) and where they do not ($p \geq 0.05$). A higher agreement rate means more reliability for t-test results under LDP. On the right plot, the grey dotted line indicates the percentage of groups below the p-value threshold ($p < 0.05$).

Likely, this is due to the t statistic being related to the sample standard deviation of the data. The standard deviation does not scale with the number of participants, since more noisy samples just make the variance increase. Figure 6 also shows that the number of random groups with $p < 0.05$ (grey dotted line) is around 70%. This implies that a sufficient number of experiments was run for both the cases $p < 0.05$ and $p \geq 0.05$.

6.3 Resilience against 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 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 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).

7 DISCUSSION

In this section, we discuss further implementation details of our proposed crowdsourcing platform.

Independent reports. Under LDP, participants are able to publish multiple independent reports, meaning that the analyst has no way to know whether two records belong to the same users. This is an intended behaviour, which allows to achieve anonymity without further distributing the privacy budget. However, submitting reports independently precludes the possibility of studying them in the temporal dimension. Moreover, although our experiments show that suitably calibrated noise can limit the error, using LDP inevitably reduces the data utility, and thus the accuracy of the results. Our recommendation is to use LDP in preliminary analyses, where the limited resources typically reduce the possibility of recruiting many participants in person.

LDP and dataset disclosure. Using LDP enables the analyst to publish the collected data, allowing others to reproduce and verify the results. Making crowdsourced data public would undoubtedly benefit the research community, and improve the credibility of studies that rely on such information. Indeed, depending on the utilized crowsourcing framework, precautions need to be taken in order to avoid accidental privacy leaks. For instance, in our crowdsourcing setting presented in section 4, the data need to be shuffled after being received by the analyst. This way, the published data will not be in the same order in which the third-party server submitted them (encrypted) to the analyst.

Plausible deniability. Besides mitigating linking attacks, LDP provides *plausible deniability* against sensitive inferences. Even if an attacker successfully manages to identify the owner of a randomized record, this will still contain imprecise information. As a consequence, the adversary gleans less information compared to what he would have obtained from an original record.

LDP in practice. Our work shows that practical application of LDP on wearable IoT records provides a much higher level of protection compared to the theoretical guarantees. Our findings match other related works on differential privacy, e.g., regarding practical

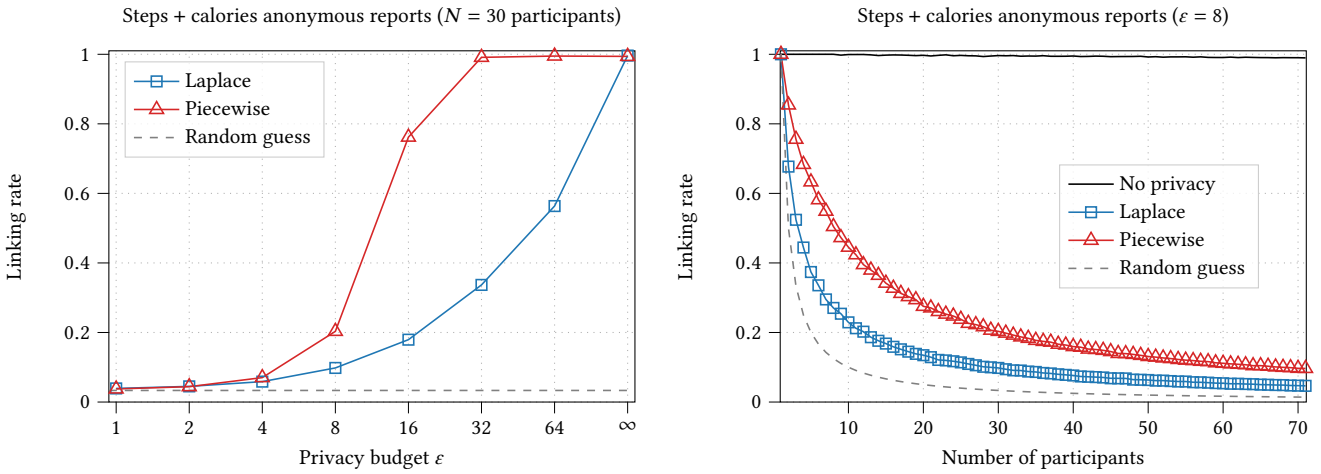

Figure 7: Linking rate for varying number of participants and privacy budget ε. A lower linking rate implies more privacy. For a same (N, ε) pair, the Laplace mechanism provide more protection against linking attacks.

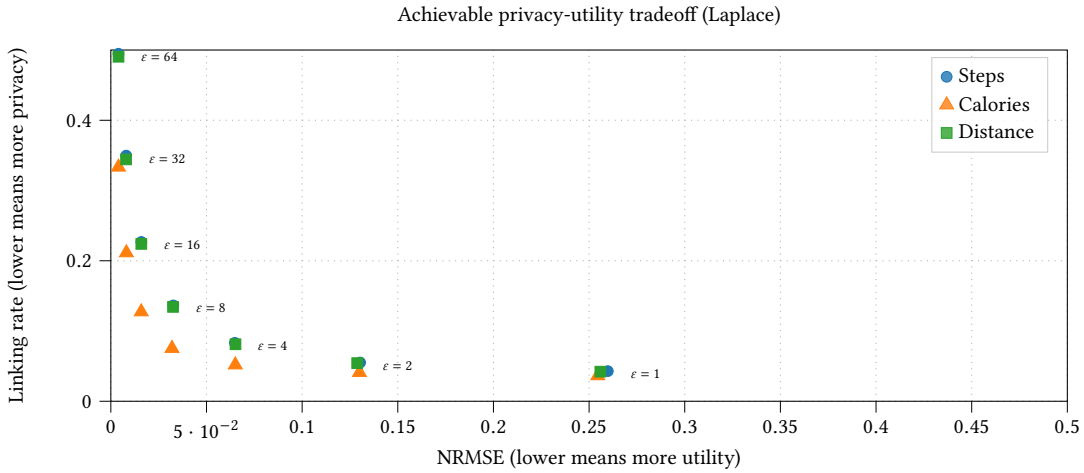

Figure 8: Privacy-utility tradeoff achieved by the Laplace mechanism for $N = 30$ participants and different values of ε. It appears that a privacy budget between 4 and 8 offers the best tradeoff.

membership inference [16] and secrets extraction [5] from machine learning models. This motivates further study of inference attacks against data protected with DP and LDP.

Report synchronization. Another detail that should be covered is the timing aspect of how the server exchanges the reports with the analyst. If the server sends the anonymous reports to the analyst as soon as they are submitted by the platform users, the analyst may be able to identify users based on when the reports are received. A more suitable strategy would be for the server to withhold the reports and forward all of them to the analyst at the end of the day.

8 CONCLUSION

The work of this paper yields encouraging results and represents a practical step towards the use of LDP in wearable IoT data crowdsourcing. Collecting independent reports as discussed in section 4 allows users to submit data over multiple days without consuming additional privacy budget. Our experiments show that LDP can indeed be used to protect individual users who share their data on crowdsourcing platforms. Furthermore, by suitably calibrating the privacy budget ($\varepsilon = 8$ for the Laplace mechanism, $\varepsilon = 4$ for the piecewise), data sanitized with LDP are still usable. When data are crowdsourced for randomized controlled trials, LDP allows to determine the statistical significance of the results in over 90% of the cases.

ACKNOWLEDGMENTS

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: RAIS – Real-time

Analytics for the Internet of Sports. The content of this paper reflects the views only of their author(s). The European Commission/ Research Executive Agency are not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Sharmin Afrose, Danfeng Daphne Yao, and Olivera Kotevska. 2021. Measurement of Local Differential Privacy Techniques for IoT-based Streaming Data. In *2021 18th International Conference on Privacy, Security and Trust (PST)*. IEEE, 1–10.

[2] Arno Appenzeller, Nick Terzer, Erik Krempel, and Jürgen Beyerer. 2022. Towards Private Medical Data Donations by Using Privacy Preserving Technologies. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*. 446–454.

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

[4] Antoine Boutet, Carole Frindel, Sébastien Gambs, Théo Jourdan, and Rosin Claude Ngueveu. 2021. 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*. 672–686.

[5] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*. 267–284.

[6] 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. 2019. Examining consumers' adoption of wearable healthcare technology: The role of health attributes. *International journal of environmental research and public health* 16, 13 (2019), 2257.

[7] Michelle M Christovich. 2016. Why Should We Care What Fitbit Shares: A Proposed Statutory Solution to Protect Sensitive Personal Fitness Information. *Hastings Comm. & Ent. LJ* 38 (2016), 91.

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

[9] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

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

[11] Fitbit. 2020. Fitbit publication and Research Library. https://healthsolutions.fitbit.com/research-library/

[12] Robert Furberg, Julia Brinton, Michael Keating, and Alexa Ortiz. 2016. *Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016.* https://doi.org/10.5281/zenodo.53894

[13] Mostafa Haghi, Kerstin Thurow, and Regina Stoll. 2017. Wearable devices in medical internet of things: scientific research and commercially available devices. *Healthcare informatics research* 23, 1 (2017), 4–15.

[14] Steven G Hershman, Brian M Bot, Anna Shcherbina, Megan Doerr, Yasbanoo Moayedi, Aleksandra Pavlovic, Daryl Waggott, Mildred K Cho, Mary E Rosenberger, William L Haskell, et al. 2019. Physical activity, sleep and cardiovascular health data for 50,000 individuals from the MyHeart Counts Study. *Scientific data* 6, 1 (2019), 1–10.

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

[16] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*. 1895–1912.

[17] British Medical Journal. 2010. Study design and choosing a statistical test. https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/13-study-design-and-choosing-statisti

[18] Andrei Kazlouski, Thomas Marchioro, and Evangelos P Markatos. 2022. What your Fitbit says about you: De-anonymizing users in lifelogging datasets.. In *SECRYPT*. 341–348.

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

[20] Hee Jin Kim, Kang Hyun Lee, Jung Hun Lee, Hyun Youk, and Hee Young Lee. 2022. The Effect of a Mobile and Wearable Device Intervention on Increased Physical Activity to Prevent Metabolic Syndrome: Observational Study. *JMIR mHealth and uHealth* 10, 2 (2022), e34059.

[21] Jong Wook Kim, Jong Hyun Lim, Su Mee Moon, Hoon Yoo, and Beakcheol Jang. 2019. Privacy-preserving data collection scheme on smartwatch platform. In *2019 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 1–4.

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

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

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

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

[26] Thomas Marchioro, Andrei Kazlouski, and Evangelos P Markatos. 2021. User Identification from Time Series of Fitness Data.. In *SECRYPT*. 806–811.

[27] OpenHumans. 2016. Open Humans Fitbit Connection. https://www.openhumans.org/activity/fitbit-connection. https://www.openhumans.org/activity/fitbit-connection

[28] Jay A Pandit, Jennifer M Radin, Giorgio Quer, and Eric J Topol. 2022. Smartphone apps in the COVID-19 pandemic. *Nature Biotechnology* 40, 7 (2022), 1013–1022.

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

[30] Yazdan Ahmad Qadri, Ali Nauman, Yousaf Bin Zikria, Athanasios V Vasilakos, and Sung Won Kim. 2020. The future of healthcare internet of things: a survey of emerging technologies. *IEEE Communications Surveys & Tutorials* 22, 2 (2020), 1121–1167.

[31] Munshi Saifuzzaman, Tajkia Nuri Ananna, Mohammad Jabed Morshed Chowdhury, Md Sadek Ferdous, and Farida Chowdhury. 2022. A systematic literature review on wearable health data publishing under differential privacy. *International Journal of Information Security* (2022), 1–26.

[32] Vajira Thambawita, Steven Alexander Hicks, Hanna Borgli, Håkon Kvale Stensland, Debesh Jha, Martin Kristoffer Svensen, Svein-Arne Pettersen, Dag Johansen, Håvard Dagenborg Johansen, Susann Dahl Pettersen, et al. 2020. Pmdata: a sports logging dataset. In *Proceedings of the 11th ACM Multimedia Systems Conference*. 231–236.

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

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

[35] Qingqing Ye, Haibo Hu, Ninghui Li, Xiaofeng Meng, Huadi Zheng, and Haotian Yan. 2021. Beyond value perturbation: Local differential privacy in the temporal setting. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10.

[36] Sofia Yfantidou, Christina Karagianni, Stefanos Efstathiou, Athena Vakali, Joao Palotti, Dimitrios Panteleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Šarūnas Girdzijauskas. 2022. LifeSnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Scientific Data* 9, 1 (31 Oct 2022), 663. https://doi.org/10.1038/s41597-022-01764-x

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

[38] Guokang Zhu, Jia Li, Zi Meng, Yi Yu, Yanan Li, Xiao Tang, Yuling Dong, Guangxin Sun, Rui Zhou, Hui Wang, et al. 2020. Learning from large-scale wearable device data for predicting epidemics trend of COVID-19. *Discrete Dynamics in Nature and Society* 2020 (2020).

A OPTIMAL LINKING CRITERIA

The maximum a posteriori probability (MAP) criterion selects the most likely sample between $y_1, \ldots, y_N$ according to

$$\hat{y} = \arg \max_{y_i, i=1,\ldots,N} \Pr[X_i = x | Y_1 = y_1, \ldots, Y_N = y_N]. \quad (24)$$

If all the samples have equal prior probability of being the anonymous report for x^*, then the MAP criterion is equivalent to the maximum likelihood (ML) criterion, i.e.,

$$\hat{y} = \arg \max_{y_i, i=1,\ldots,N} \Pr[Y_1 = y_1, \ldots, Y_N = y_N | X_i = x^*]. \quad (25)$$

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:

$$\hat{y} = \arg \max_{y_i, i=1,\ldots,N} p(y_i|x^*) = \arg \max_{y_i, i=1,\ldots,N} \prod_{f=1}^{m} p(y_i[f]|x^*[f]), \tag{26}$$

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 $i = 1, \ldots, N$. 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

$$\hat{y} = \arg \max_{y_i, i=1,\ldots,N} \prod_{f=1}^{m} \exp(-\varepsilon \frac{|y_i[f] - x^*[f]|}{\Delta[f]}) \tag{27}$$

$$= \arg \min_{y_i, i=1,\ldots,N} \sum_{f=1}^{m} \frac{|y_i[f] - x^*[f]|}{x_{\max}[f] - x_{\min}[f]}, \tag{28}$$

which results in our criterion.

For the Piecewise mechanism, instead,

$$\hat{y} = \arg \max_{y_i, i=1,\ldots,N} \prod_{f=1}^{m} \begin{cases} e^{\varepsilon} & \text{if } y_i[f] \in (L(x[f]), R(x[f])) \\ 1 & \text{if } y_i[f] \notin (L(x[f]), R(x[f])) \end{cases}, \tag{29}$$

$$= \arg \max_{y_i, i=1,\ldots,N} \sum_{f=1}^{m} \chi \{y_i[f] \in (L(x[f]), R(x[f]))\} \tag{30}$$

B BOUNDS ON THE LINKING RATE

In order to find a bound to the success probability of a linking attack, we analyze the complementary event $\mathcal{S}^c$. Indeed, complementary events are linked by $\Pr[\mathcal{S}] + \Pr[\mathcal{S}^c] = 1$, even when they are conditioned. 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. 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 $\Pr[\mathcal{S}^c|1, N, \varepsilon, \text{Lap}] = \eta$

$$\eta \geq \Pr[Y_i \notin (x^* - \Delta, x^* + \Delta) \wedge Y_j \in (x^* - \Delta, x^* + \Delta), j \neq i] \tag{31}$$

$$= \Pr[Y_i \notin (x^* - \Delta, x^* + \Delta)] \left(1 - \prod_{j=1, j\neq i}^{N} \Pr[Y_j \notin (x^* - \Delta, x^* + \Delta)]\right) \tag{32}$$

$$\geq e^{-\varepsilon} \left(1 - \left(\frac{1}{2} + \frac{1}{2} e^{-2\varepsilon}\right)^{N-1}\right). \tag{33}$$

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 $f = 1, \ldots, m$ (and conversely, all the features of another report fall within the region). Furthermore, each of the m features is randomized with privacy budget ε/m. The bounds, thus, becomes

$$\Pr[\mathcal{S}^c|m, N, \varepsilon, \text{Lap}] \geq e^{-\varepsilon} \left(1 - \left(1 - \left(\frac{1}{2} - \frac{1}{2} e^{-\frac{2\varepsilon}{m}}\right)^{m}\right)^{N-1}\right). \tag{34}$$

which leads to eq. 22.

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 $\Pr[\mathcal{S}^c|1, N, \varepsilon, \text{PW}] = \eta'$,

$$\eta' \geq \Pr[Y_i \notin (L(x^*), R(x^*)) \wedge Y_j \in L(x^*), R(x^*)), j \neq i] \tag{35}$$

$$= \Pr[Y_i \notin (L(x^*), R(x^*))] \left(1 - \prod_{j=1, j\neq i}^{N} \Pr[Y_j \notin (L(x^*), R(x^*))]\right) \tag{36}$$

$$\geq \frac{\tau}{\tau + e^{\varepsilon}} \left(1 - \left(1 - \left(\frac{\tau + e^{\varepsilon} - 1}{\tau + e^{\varepsilon}}\right)^{N-1}\right)\right) \tag{37}$$

$$= \frac{e^{\varepsilon/3}}{e^{\varepsilon/3} + e^{\varepsilon}} \left(1 - \left(1 - \left(\frac{e^{\varepsilon/3} + e^{\varepsilon} - 1}{e^{\varepsilon/3} + e^{\varepsilon}}\right)^{N-1}\right)\right) \tag{38}$$

Repeating the same reasoning for multiple features, we get

$$\Pr[\mathcal{S}^c|1, N, \varepsilon, \text{PW}] \geq \frac{e^{\frac{\varepsilon}{3}}}{(e^{\frac{\varepsilon}{3m}} + e^{\frac{\varepsilon}{m}})m} \left(1 - \left(1 - \frac{1}{(e^{\frac{\varepsilon}{3m}} + e^{\frac{\varepsilon}{m}})m}\right)^{N-1}\right), \tag{39}$$

which leads to eq. 23. Notably, this bound holds also if the attacker adopts the optimal strategy from eq. 21.