Interval Privacy: A Framework for Privacy-Preserving Data Collection

Jie Ding and Bangjun Ding

Abstract

The emerging public awareness and government regulations of data privacy motivate new paradigms of collecting and analyzing data transparent and acceptable to data owners. We present a new concept of privacy and corresponding data formats, mechanisms, and theories for privatizing data during data collection. The privacy, named Interval Privacy, enforces the raw data conditional distribution on the privatized data to be the same as its unconditional distribution over a nontrivial support set. Correspondingly, the proposed privacy mechanism will record each data value as a random interval (or, more generally, a range) containing it. The proposed interval privacy mechanisms can be easily deployed through survey-based data collection interfaces, e.g., by asking a respondent whether its data value is within a randomly generated range. Another unique feature of interval mechanisms is that they obfuscate the truth but not perturb it. Using narrowed range to convey information is complementary to the popular paradigm of perturbing data. Also, the interval mechanisms can generate progressively refined information at the discretion of individuals, naturally leading to privacy-adaptive data collection. We develop different aspects of theory such as composition, robustness, distribution estimation, and regression learning from interval-valued data. Interval privacy provides a new perspective of human-centric data privacy where individuals have a perceptible, transparent, and simple way of sharing sensitive data.

Index Terms

data collection, human-computer interface, interval privacy, interval mechanism, privacy, survey.

J. Ding is with the School of Statistics, University of Minnesota Twin Cities, MN 55414, USA. B. Ding is with the School of Statistics and Finance, East China Normal University, Shanghai, China.
CONTENTS

I Introduction

II Interval Privacy
   II-A Notation .................................................. 7
   II-B Interval Privacy .......................................... 8
   II-C Explanations of Interval Privacy via Simple Examples ............... 9

III General Interval Mechanisms, Data Formats, and Theoretical Foundations 15
   III-A Canonical Interval Mechanism .............................. 15
   III-B Fundamental Properties of Interval Mechanism .................... 18
   III-C Extension: Interval Mechanism of General Topology ............... 20
   III-D Extension: Individual-Level Privacy Enhancement .................. 23
   III-E Regression with Private Responses ............................. 26

IV Experiments
   IV-A Estimation of Moments ....................................... 28
   IV-B Estimation of Regression Functions ............................ 29
   IV-C Tradeoff between Learning and Privacy Coverage .................. 30
   IV-D Sensitivity of Misspecified Noise .............................. 30
   IV-E Case Study: Distribution from Individual-Adaptive Surveys ........ 32
   IV-F Case Study: Life Expectancy Regression ........................ 34

V Related Literature

VI Conclusion and Further Remarks

References
I. Introduction

With new and far-reaching laws such as the General Data Protection Regulation [1] and frequent headlines of large-scale data breaches, there has been a growing societal concern about how personal data are collected and used [2], [3]. Consequently, data privacy has been an increasingly important factor in designing signal processing and machine learning services [4], [5]. This paper will address the following scenario often seen in practice. Suppose that Alice is the agent who creates and holds raw data, which will be collected by another agent Bob. On the one hand, Alice may not trust Bob or the transmission channel to Bob. On the other hand, Bob is interested in population-wide inference using the data provided by Alice and many other individuals, but not necessarily the exact value held by Alice.

The above learning scenario is quite common in, e.g., Machine-Learning-as-a-Service cloud services [6], [7], multi-organizational Assisted Learning [8], [9], survey-based inferences [10], [11], and information fusion [12], [13]. The formalization of individual-level data privacy and population-level estimation utility has motivated active research on what is generally referred to as local data privacy across fields such as data mining [14], security [15], and statistics [16], and information theory [17]. The general goal of local data privacy is to suitably process raw data (often by randomizations) during the data collection and evaluate it through an appropriate framework.

In this work, we propose a notion of local privacy named interval privacy for protecting data collected for further inferences. The main idea is to enforce privacy in such a way that the distribution of the raw data conditional on its privatized data remains the same (up to a normalizing constant) on a moderately large support set. In other words, no additional information is gained except that the support of the data becomes narrow. Accompanying the notion of interval privacy, we use the size (in a measure-theoretic sense) of the conditional support to quantify the level of privacy. The size, named as privacy coverage, enables a natural interpretation and perception of the amount of ambiguity exposed to the data collector. We then introduce interval privacy mechanisms for realizing data collection in practice.

Our perspective of privacy is motivated by the following practical concerns. Suppose that an organization collects privacy-sensitive information from individuals, e.g., an organization gathers users’ demographic information. A concern is how to develop a data collection interface so that individuals can easily perceive that the collected data are at their discretion. In other words,
individuals do not have to submit exact raw data first and then rely on any subsequent processing of those data, which can be a black-box procedure obscure for the public. The main idea of interval mechanism is to generate random intervals that partition the data domain and collect the interval containing the underlying data. It can be naturally implemented as a transparent yet simple survey interface, where an individual can directly see the ultimate collection and perceive its ambiguity. As individuals may have different privacy sensitivities, another related concern is how to obtain data in a way adaptive to individual-level privacy. Interval privacy addresses this by progressively collecting data from wider (and thus more private) intervals to narrower ones, meaning that individuals may respond, not respond, or respond further, at their discretion.

Our notion of privacy naturally leads to a new form of disclosing and collecting sensitive data, namely representing them as intervals instead of points. For example, a sensor’s accurate distance with the target $y = 10$ (in meters) is privatized by first generating a random threshold, say 20, and then publicizing the corresponding interval $(-\infty, 20]$; or an individual’s 60k salary (in dollars) is privatized by first generating random thresholds, say 41k and 85k, and then reporting the interval $(41k, 85k]$. The random thresholds can be generated from any distribution known to the data collector, e.g., Gaussian, Logistic, and Uniform distributions, independent of the underlying data. It is worth noting that an interval is not necessarily symmetric around the underlying raw data value. Tab. I illustrates $Y$ and its private counterpart in a dataset that we will revisit in experimental studies. Each individual’s privacy coverage describes the interval size or level of ambiguity. For example, the data with 97% coverage is less private than the 99% one, which is in line with the perception that $Y < 82.2$ reveals more information than $Y < 85.4$ does. We will show several fundamental properties of the proposed privacy mechanism to render its broad applicability. These include the composition property that characterizes the level of overall privacy degradation in the presence of multiple queries to the same data, robustness to pre-processing, robustness to post-processing, distributional identifiability, and extensions from intervals to general ranges. We will demonstrate the use of interval-private data for several inference tasks, including moment estimation, functional estimation, and supervised regression. For the particular case of supervised regression where the response labels are privatized, estimating the data-generating function is highly nontrivial when the number of predictors is large. We propose an algorithm to estimate regression functions using the interval privatized data.

The main contributions of this paper are summarized below.

- We develop a new perspective of data privacy named interval privacy, particularly suitable
TABLE I
A snapshot of the ‘life expectancy’ database [18] to be studied in Subsection IV-E. The life expectancy, \( Y \), is privatized into random intervals, with an overall privacy coverage of 60.3%. 

| \( Y \) (in years) | 59.3 | 82.3 | 79.5 | 51.7 |
|-------------------|------|------|------|------|
| Privatized \( Y \) | (0, 82.2] | (0, 85.4] | (64.2, \( \infty \)) | (46.9, 72] |
| Privacy coverage  | 97%  | 99%  | 70%  | 47%  |

Fig. 1. A generic data collection system based on interval privacy.

for privacy-sensitive data collection. We develop interval mechanisms and show their desirable interpretations to implement interval privacy naturally. Fig. 1 illustrates a general use scenario, where individuals’ private data are collected through an interface that masks each data point into an interval (or, in general, a range). We show several unique features of an interval privacy mechanism. First, it tells the truth while obfuscating the truth, which is essential for application domains requiring information fidelity (such as census and defense scenarios). Second, it can be easily deployed through survey systems, with an interpretable and perceptible human-computer interface. Third, such an interface can allow progressive narrowing of collected intervals and thus be adaptive to individuals’ privacy sensitivities that are likely to vary in practice. Fig. 2 illustrates a general survey system built upon interval privacy, which, unlike conventional surveys widely used in many fields such as sociology, political science, and psychometrics [10], [11], [19], generates questions in an individual-specific and data-adaptive manner. To our best knowledge, this is the first work
that advocates the use of random ranges for privacy-preserving data collection and the use of randomly generated questions in survey design.

- We develop fundamental properties of the proposed privacy mechanism, including the composition property that characterizes privacy leakage under multiple queries of the same data, the robustness to pre-processing and post-processing, and the identifiability of the underlying data distributions. We exemplify the use of interval privacy in estimating population distribution, statistical functional, and regression function, and show that the data collector does not necessarily need to know the distributional form of raw data for accurate population-level inference. In particular, we provide a general theory to show the topology and probabilistic structures needed to reconstruct the population distribution from random ranges non-parametrically. We develop several extensions to address individual-level privacy guarantees. We also develop a general method to perform supervised regression with interval-privatized responses. The technique can be applied to various interval-private data types, including pure intervals or a mixture of intervals and points.

- We experimentally demonstrate the proposed concepts, data formats, properties, and methods. We also discuss the connections between interval privacy and the existing literature from multiple angles. For example, we will point out (in the Appendix) that interval privacy is neither a generalization nor a specialization of local differential privacy [14], [15], a popular framework that has been extensively studied.

The rest of the paper is outlined below. In Section II we introduce the basic concept of interval privacy and use simple examples to explain its key ideas and use scenarios. In Section III we introduce general interval privacy mechanisms, data formats, theoretical properties, and various practical implications. In particular, Subsection III-A introduces a canonical form of mechanisms, Subsection III-B establishes several properties of the mechanisms, Subsection III-C extends the scope of intervals to general ranges, Subsection III-D extends the mechanisms to address individual-level privacy guarantee and individual-adaptive data collection. In Subsection III-E, we present an approach to use the interval data for regression learning. In Section IV we provide some experimental studies. We conclude the paper in Section VI. The Appendix includes technical details and additional discussions.
II. INTERVAL PRIVACY

A. Notation

We let $Y$ denote a continuously-valued random variable representing the raw data throughout the paper. Suppose that the raw data $Y_1, \ldots, Y_n \in \mathcal{Y} \subset \mathbb{R}$ are i.i.d. with probability $P_Y$, density $p_Y$, and cumulative distribution function (CDF) $F_Y$. We will write $i = 1, \ldots, n$ as $i \in [1 : n]$. For a random vector $Q$, we let $Q^{(i)}$ and $Q_i$ denote its $i$-th entry and $i$-th observation, respectively, unless otherwise stated.

We consider the local data privacy scenario where there are many data owners and one data collector. A data owner is an individual that holds a private data value (represented by $Y$), who does not trust the data collector. A data collector’s genuine goal is to infer distributional information of $Y$ instead of each individual’s data value. As such, a general local privacy scheme uses a random mechanism $\mathcal{M}$ that maps each $Y$ to another variable $Z \in \mathcal{Z}$ and then collects $Z$. The mechanism is often represented by a conditional distribution of $Z \mid Y$. The random variable $Z$ may be constructed by a measurable function of $Y$, or a function of $Y$ and other auxiliary parameters.

Fig. 2. Illustration of a general survey system that generates different human-computer interfaces for participating individuals.
random variables. We assume that the joint distribution of \([Y, Z]\) exists and has a density with respect to the Lebesgue measure. Suppose that \(S\) is a Borel set. We let \(L(S) \triangleq \mathbb{P}_Y(S)\) denote the ‘size’ of \(S\) (which remains the same throughout the paper).

For two sets \(A, B\), let \(A - B\) denote the set of elements that belong to \(A\) but not belong to \(B\). Let \(o_p(1)\) denote a sequence of random variables that converges to zero in probability.

**B. Interval Privacy**

**Definition 1 (Interval Privacy).** A mechanism \(\mathcal{M}\) has the property of interval privacy if almost surely for all \(y_1, y_2 \in S_z\),

\[
\frac{p_{Y \mid Z}(y_1 \mid Z = z)}{p_{Y \mid Z}(y_2 \mid Z = z)} = \frac{p_Y(y_1)}{p_Y(y_2)},
\]

(1)

where \(p_{Y \mid Z}\) denotes the distribution of \(Y\) conditional on \(Z\) and \(S_z\) is the support of \(Y\) given \(Z = z\).

The privacy coverage of \(\mathcal{M}\), denoted by \(\tau(\mathcal{M})\), is defined by \(\mathbb{E}(L(S_Z))\), where the expectation is over \(Z\), and \(L\) is the size under the prior distribution of \(Y\) (namely \(p_Y\)). An \(\mathcal{M}\) is said to have \(\tau\)-interval privacy if \(\tau(\mathcal{M}) \geq \tau\).

**Implication 1:** We will give explicit formulas for some particular designs in Section II-C.

Equation (1) means that the conditioning on \(Z = z\) does not provide extra information except that \(y\) falls into \(S_z\). If \(y_1 \neq y_2\), and they fall into the same support \(S_z\), then their likelihood ratio remains the same as if no action were taken. Equation (1) also implies that

\[
p_{Y \mid Z}(y \mid Z = z) = c_z \mathbb{1}_{y \in S_z} \cdot p_Y(y)
\]

(2)

holds for the normalizing constant \(c_z = 1/\int_{S_z} p_Y(y)dy\).

**Implication 2:** Suppose that \(Y = y_1\) is to be protected. Interval privacy creates ambiguity by obfuscating the observer with sufficiently many \(y_2\)’s in a neighborhood whose posterior ratios do not vary by incorporating the new information \(Z = z\). Also, suppose that \(S_z\) is a (closed or open) interval, then the finite cover theorem implies the following alternative to the above second condition. For all \(y\) in the interior of \(S_z\), there exists an open neighborhood of \(y\), \(U(y) \subset S_z\), where (1) holds for all \(y_1, y_2 \in U(y)\).

**Implication 3:** By its definition, the privacy coverage \(\tau(\mathcal{M})\) takes values from \([0, 1]\). The privacy coverage quantifies the average amount of ambiguity or the level of privacy. A larger value indicates increased privacy. Likewise, for each (nonrandom) raw-privatized data pair, \((y, z)\),
we introduce \( L(S_z) \) as the *individual privacy coverage*, interpreted as the privacy level for a particular data item being collected (illustrated in the third row of Tab. 1).

A related measure is \( 1 - \tau(M) \) which naturally describes the *privacy leakage*. For instance, the coverage of \( \mathcal{Y} \) is one, and the leakage is zero, meaning no privacy is leaked; Meanwhile, the coverage of \( y \) (as a degenerate interval) is zero. To realize interval privacy, we will introduce natural interval mechanisms that convert \( y \) to a random interval that contains \( y \). For example, \( S_z \) is in the form of \((-\infty, u]\) or \((u, \infty)\), encoded by the vector \( z = [u, 1_{y \leq u}] \).

We provide Fig. 3 to visualize our unique approach to protecting data information. It shows the data format of interval data and its released information of the raw data as implied by posterior uncertainty. It also visualizes the popular approach that privatizes data by perturbations. From a Bayesian perspective, the perturbation changes the density shape, while the interval approach changes the essential support. In the plot, we generated raw data \( y \) from the standard Gaussian. The interval privacy used the standard Logistic random variable \( U \) and reports either ‘\( \leq U \)’ or ‘\( > U \)’, resulting in around 0.25 privacy leakage. The perturbation approach truncated the raw data within \([-3, 3]\) and added the Laplacian noise so that it achieves a 2-local differential privacy.

C. Explanations of Interval Privacy via Simple Examples

This section provides simple examples of data formats, mechanisms, and practical implications regarding interval privacy. We will introduce formal definitions of different mechanisms and theoretical foundations in Section III.

Suppose that a data analyst aims to study the population distribution of salary. To collect the salary information from an individual (say Alice) without revealing the underlying value, Alice is asked to report whether the salary is above a threshold or not. This is illustrated in Fig. 4(a). This naturally leads to the following privacy mechanism, perhaps the simplest interval mechanism. Only the indicator of whether the salary is larger than a randomly generated threshold is reported.

- **Case-I interval mechanism**: Let \( U \in \mathbb{R} \) be a random variable independent with \( Y \), referred to as an anchor point. Either \( Y \leq U \) or \( Y > U \) is observed. The observations are \( n \) i.i.d. copies of \( Z = [U, \Delta] \), where \( \Delta = 1_{Y \leq U} \) is an indicator variable.

Likewise, we also define the following mechanism that admits a bounded interval (e.g., \$60-80k\). More general mechanisms will be introduced in Section III The names of ‘Case-I’ and ‘Case-II’ follow the convention of the same types of data historically studied in survival analysis [20].
Fig. 3. An illustration of interval privacy (left column) and local differential privacy with noise perturbation (right column) in terms of: the raw data $y$ and interval-privatized data (left-up), the posterior distribution of $Y$ given an interval $[-1,2]$ (left-bottom), raw data $y$ and Laplacian-perturbed (point) data (right-up), and the posterior of $Y$ given an observed point 2 (right-bottom).

- **Case-II interval mechanism:** Let $[U, V] \in \mathbb{R}^2$ be a random variable that satisfies $P(U \leq V) = 1$ and is independent with $Y$. Either $Y \leq U$, $U < Y \leq V$, or $Y > V$ is observed. The observations are $n$ i.i.d. copies of $Z = [U, V, \Delta, \Gamma]$, where $\Delta = 1_{Y \leq U}$ and $\Gamma = 1_{U < Y \leq V}$ are indicator variables.

We summarize some key characteristics of interval privacy mechanisms below.

1) **Conditional non-informativeness:** We will show in Subsection III-A that the collected data $Z$ in the above examples satisfy the interval privacy (Definition 1). Thus, conditional on the revealed support set, e.g., $(-\infty, U]$, no additional information is revealed since the relative probability densities of $Y$ conditional on $Y \leq U$ do not differ from unconditional ones. We also refer to this property as conditional non-informativeness, meaning that the only information
provided by the collected data $Z$ about the raw data $Y$ is an (often wide) range that contains $Y$.

2) **Information fidelity**: An interesting aspect of the interval privacy mechanism is that it collects masked data instead of perturbed data. This offers two practical benefits. First, maintaining information fidelity is vital in many applications domains such as census, security, and defense, where collecting perturbed data can lead to misinterpretations or disastrous decisions. The interval-private data convey information without lying about the underlying values. As we will show later, even if each point is masked into a fairly wide range, one can still reconstruct the underlying population distribution without systematic biases (asymptotically).

Second, consider many scenarios where a resourceful organization collects private information from anonymized individuals. Individual hope to easily perceive that the already-collected data are indeed private. Existing privacy schemes such as homomorphic encryption [21] and local differential privacy [14] often need the collecting organization to implement sophisticated cryptography- or randomization-based procedures at the backend. Consequently, their privacy architectures require individuals to submit exact raw data in the collecting interface, which inevitably raises trustworthiness issues. A potential remedy is immediately applying privatization after collecting an individual’s data and publicizing the implementation source codes. But even in that case, it may not be fully transparent to individuals (especially to the public). In contrast, the proposed interval privacy mechanisms can be transparently deployed through electronic survey-based data collection infrastructures. Such a privacy interface allows each individual to perceive
3) **Distributional identifiability**: It is worth noting that generating random $U$ in the above Case-I mechanism is essential. If $U$ is deterministic, it is impossible to accurately estimate the distribution of $Y$ since one can always find a distinct distribution whose mass on the pre-determined intervals coincide. Suppose that the essential support of $U$ contains that of $Y$. It has been shown under reasonable conditions that the distribution of $Y$ can be consistently estimated from interval observations even if the underlying distribution is not parameterized [22]. We will revisit the nonparametric estimation method and develop new theories for general interval mechanisms in Subsection III-C. To illustrate distributional identifiability, we generate 1000 points of $Y$ from a standard Logistic distribution, and Case-I interval-private data from $U$ that follows a Logistic distribution whose scale is 2. Fig. 5 (left plot) shows parametric and nonparametric estimations of the CDF $F_Y$ from the interval data. The parametric estimation is based on the standard maximum likelihood approach. The nonparametric estimation uses the self-consistency algorithm [23] implemented in the ‘Icens’ R package [24].

**Example 1 (Functional Estimation).** Suppose that an analyst is interested in estimating a smooth functional $K(F_Y)$ of the underlying distribution function $F_Y$. A nonparametric estimator is $\hat{K}(\hat{F}_Y)$ where $\hat{F}_Y$ is the nonparametric maximum likelihood estimator of $F_Y$ [22]. Specifically, all moment functionals $K : F_Y \mapsto \int Y y^k dF_Y(y)$, or more generally, linear functionals in the form of $K : F_Y \mapsto \int Y \phi(y) dF_Y(y)$ can be estimated in this way.

**Example 2 (Mean Estimation).** Sometimes, a statistical functional may be directly estimated without the need of estimating $F_Y$. For example, suppose that the raw data are i.i.d. $Y_i \in [a,b]$ for $i \in [1:n]$, with unknown mean $\mu$. The observations are $Z_i = [U_i, \Delta_i]$, $i \in [1:n]$, from the Case-I mechanism with $U_i \sim_{i.i.d.}$ Uniform$[a,b]$. We provide the following estimator and will show that it is a consistent and unbiased estimator of $\mu$.

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \left( \Delta_i(2U_i - b) + (1 - \Delta_i)(2U_i - a) \right).$$

(3)

**Proposition 1.** The estimator in Example 2 satisfies $\mathbb{E}(\hat{\mu}_n) = \mu$ and $\text{var}(\hat{\mu}_n) = O(n^{-1})$.

4) **Achievability**: The ambiguity as quantified by privacy coverage $\tau$ (in Definition 1) can be controlled by the distribution of $[U, V]$, or the number of intervals, e.g., two in Case-I and three in Case-II. The larger $\tau$, the more ambiguity and thus more protection. The following result
shows that any privacy coverage in $[c, 1]$ for a constant $c \in [0, 1)$ is achievable by a suitable choice of $U, V$.

**Theorem 1 (Achievability).** Assume that the density function of $Y$ is bounded. For any $\tau \in (1/2, 1)$ (respectively $(1/3, 1)$) there exists a Case-I (respectively Case-II) mechanism $\mathcal{M}$ whose privacy coverage is exactly $\tau$.

The result implies that a privacy mechanism exists for arbitrarily close to one privacy coverage. Also, the proof indicates that the choice is not unique. As a by-product of the proof, $\tau(\mathcal{M}) = n^{-1} \sum_{i=1}^{n} [F_Y(u_i)^2 + (1 - F_Y(u_i))^2]$ is a consistent estimator of the privacy coverage for Case-I mechanisms. The estimator can be similarly extended for other mechanisms. Although the above result indicates that the max privacy near one is achievable, we may not do so in practice since there is an inherent tradeoff between privacy and estimation accuracy. To see that, we provide an example inference task below, which is interesting in its own right.

For any functional that is differentiable along Hellinger differentiable paths of distributions (e.g., linear functionals), one can derive an Hájek-LeCam convolution theorem type information lower bound, giving the best possible limit variance that can be attained under $\sqrt{n}$ convergence rate where $n$ denotes the data size [25]. The distribution of anchor points is said to be optimal if such information lower bound is attained by the produced interval data.

**Theorem 2 (Optimal Anchor).** An optimal distribution of $U$ (in the Case-I mechanism) for estimating any linear functional in Example 7 exists, and it has the density

$$g_U(u) = c_\phi \left\{ F_Y(u)(1 - F_Y(u)) \right\}^{1/2} |d/d_u \phi(u)|$$

if it is integrable, where $c_\phi$ is a normalizing constant.

The above result indicates a tradeoff between privacy coverage and statistical efficiency (in inference). Fig. 5 exemplifies the estimation of $F_Y$ and optimal Case-I interval mechanisms.

5) **Privacy guarantee:** In the above discussion of achievability and additional properties to be introduced in Subsection III-B, we use the privacy coverage $\mathbb{E}(L(S_Z))$ to quantify the privacy of a mechanism. A skeptical reader may ask how to ensure individual-level privacy. Recall that in Subsection II-B we introduced the individual privacy coverage $L(S_z)$, namely the size of an interval represented by $z$, to quantify an individual’s privacy. We provide two general ways to enhance individual-level privacy. Suppose that an individual has an associated “bottom line”
Fig. 5. An illustration of population inference from interval-private data. The left plot shows the true CDF of the standard Logistic random variable $Y$, and its estimation using both nonparametric and parametric methods with 1000 data. The right plot shows the density of $Y$ and optimal densities of its Case-I anchor ($U$) for estimating the first & second moments (using Theorem 2).

$\tau \in [0, 1]$, a value such that an organization can only collect an $z$ if $L(S_z) \geq \tau$. The first method uses an interval mechanism where each generated interval has coverage of at least $\tau$. Though simple, such a mechanism may not exist for some $\tau$ (e.g., $\tau = 0.6$) since we cannot have two intervals whose sizes are both at least 0.6. Moreover, the simple Case-I&II mechanisms cannot simultaneously guarantee individual-level privacy and distributional identifiability, and thus a more general topology (of data ranges) is required in the mechanism design. More on this will be discussed in Subsection III-C.

The second method simply lets an individual decide whether to report the associated interval or not, depending on the $\tau$. In practice, this can be implemented in a way illustrated in Fig. 4(b), which provides an “Not wish to answer” option. Meanwhile, the interval mechanism needs to randomly subsample reported intervals to avoid the inference of the unreported interval
(especially for a large $\tau$). Although such a mechanism introduces a selective bias (towards large intervals), we will show that the above appealing properties (such as non-informativeness and distributional identifiability) can still hold. More technical discussions will be elaborated in Subsection [III-D]

6) **Adaptivity to individual-level privacy**: The interval mechanism can be extended to a progressive fashion. We illustrate this point in Fig. [4](c). Suppose that for the first question, an individual chooses $\leq U_1$; our interface then generates the second question with $U_2 < U_1$; if the individual chooses $> U_2$, the interval $(U_2, U_1]$ is then collected. Such a progressive mechanism aligns with the above discussion of point (5), where the idea is to respect each individual’s privacy while exploiting heterogeneous privacy sensitivities. We will revisit this idea in Subsections [III-D] and [IV-E]

III. **General Interval Mechanisms, Data Formats, and Theoretical Foundations**

With the high-level explanation in Subsection [II-C] we now introduce general interval mechanisms and technical details.

A. **Canonical Interval Mechanism**

Recall that in local privacy settings, a benign data collector is only interested in the population instead of individual-level information. Each individual who holds raw data (‘$Y$’) will use a data collection interface to submit privatized data (‘$Z$’) to the collector. Our interval privacy mechanism does not collect $Y$ itself, but a privatized data $Z$ motivated by the scenarios where an individual will

- report an interval that contains $Y$,
- report $Y$ if it falls into an ‘acceptable’ range, and
- have an acceptable range independent of $Y$.

A natural mechanism to realize interval privacy is randomly partitioning the data domain $\mathcal{Y} \subseteq \mathbb{R}$ into disjoint intervals for each data owner and collecting the interval into which $Y$ falls. As such, we can naturally implement the mechanism through multi-choice survey questions, where each interval corresponds to a choice. This is illustrated in Fig. [5](a)(b). Unlike existing survey systems, our proposed system generates random (and thus different) choices for respondents. The randomness is needed to nonparametrically reconstruct the unknown population distribution
from collected data, which will be elaborated in Subsection III-C. Occasionally, the underlying point $Y$ is reported if it falls into a range that the data owner considers as non-sensitive. This can be implemented by an optional text box in the above survey, as shown in Fig. 6(b). Consequently, the collected data are in the form of intervals or a mixture of intervals and points.

Formally, we introduce the following notions. Let $Q = [Q^{(1)}, \ldots, Q^{(m-1)}]$ be a random vector with $\text{essinf } Y = Q^{(0)} = Q^{(1)} < \cdots < Q^{(m)} = \text{esssup } Y$, as illustrated in Fig. 6(a). We will refer to each $Q^{(i)}$ as an ‘anchor’ point, and let $R^{(i)} \triangleq (Q^{(i-1)}, Q^{(i)})$ for $i \in [1 : m]$. Then, the interval $R^{(i)}$ into which $Y$ falls is collected. Suppose that when $Y$ falls into a pre-determined set $A \subseteq \mathbb{R}$, named an ‘acceptable range,’ then the data owner chooses to disclose the value of $Y$. In practice, the acceptable range is at the data owner’s discretion, and the set may not be fixed. To model the real-world complexity, we suppose that $A$ can be one of the following: $\emptyset$, a fixed set, or the union of $(Q^{(k-1)}, Q^{(k)})$ for a fixed set of $k$. We suppose that the form of $A$ is pre-specified and independent with $Y$.

**Definition 2 (Canonical Interval Mechanism).** A privacy mechanism, denoted by $\mathcal{M} : Y \mapsto Z$, maps $Y$ to

$$Z = [Q, I(Q, Y), Y \cdot 1_{Y \in A}],$$

where $Q = [Q^{(1)}, \ldots, Q^{(m-1)}] \in \mathbb{R}^{m-1}$ is a random vector independent with $Y$, and $I : (Q, Y) \mapsto i$ is the indicator function defined by $Y$ falling into $(Q^{(i-1)}, Q^{(i)})$, $i \in [1 : m]$. The corresponding privacy coverage and privacy leakage follow Definition 1.

**Theorem 3 (Validity).** A mechanism $\mathcal{M}$ in Definition 2 satisfies the interval privacy in Definition 1.
Remark 1 (Interpretation of Theorem 3). Intuitively, the validity is because $I(Q,Y)$, the informative part of $Z$, only reveals the range information regarding $Y$ but not any distributional information within that range. It also allows the range to degenerate to a point when $\mathbb{1}_{Y \in A} = 1$ (if $A$ is not empty). Thus, the posterior density ratio equals the prior density ratio up to a range, as shown in (1). Also, we point out that interval mechanisms are adaptive to an individual user’s privacy preference, meaning that progressively refined information can be obtained at the discretion of individual respondents without violating Definition 7. Formally, suppose that the system also generates a second mechanism $\mathcal{M}' : Y \mapsto Z'$ based on anchor points $Q'$ that are (adaptively) supported on the inferred range from a previous mechanism $\mathcal{M}$. Then, the joint of these two mechanisms is a mechanism that meets interval privacy. This observation can be proved similarly to Theorem 3. A practical implication is that respondents may choose to answer zero, one, or more times of randomly generated surveys depending on their earlier answers and privacy preference. This point will be revisited in Example 4.

Remark 2 (Practical Implementation). In practice, a privacy-preserving data collection system involves two parties, a data owner (‘Alice’) and a data collector (‘Bob’). A general collection procedure is outlined as follows. First, a system designer, who may or may not be one of the two parties, define a way of generating $Q$. Such a generating process may be open-source implemented so that it is transparent to both parties. Second, the two parties agree on the use of the mechanism for data collection. Third, for Alice’s data value $Y$, an instance of $Q$ is generated, and Alice reports the interval to Bob. Additionally, Alice has the option to report the exact value, but this is at Alice’s discretion. In the end, the set of data Bob collects consists of intervals and possibly some exact values (degenerate intervals).

Remark 3 (Interpretation of Data). The observables include a partition of $\mathcal{Y}$ (by $Q$), the interval that $Y$ falls (by $I(Q,Y)$), and sometimes the value of $Y$ (represented by $Y \mathbb{1}_{Y \in A}$). The information obtained from the privatized data $Z$ is an interval containing $y$. It can be seen that an interval mechanism does not lie about the value of $y$. This property does not hold for popular approaches where perturbations are injected into the raw data.

An alternative notation to $I(Q,Y)$ is to use $m$ indicator variables $\mathbb{1}_{Y \leq Q(1)}, \ldots, \mathbb{1}_{Y \leq Q(m)}$ to represent where $Y$ is located at. By the definition, $\mathbb{1}_{Y \leq Q(i)} = 1$ if $i \leq I(Q,Y)$ and $\mathbb{1}_{Y \leq Q(i)} = 0$ otherwise. The values of $Q$ are random so that it is possible to identify the population distribution of $Y$ (elaborated in Subsection III-C). So then, the randomness of $Z$ conditional on $Y$ comes
from $Q$. The choice of $Q$ determines privacy-utility tradeoffs. Consider an extreme case where $m$ is sufficiently large. Then, the collected interval tends to be narrow, and the privacy coverage tends to zero. In another case where $m = 1$ and $Q \in \mathbb{R}$ has a considerable variance, the interval is likely to be close to $(-\infty, \infty)$, which enjoys good privacy but offers little utility in distribution estimation.

**Remark 4 (Interpretation of $A$).** The acceptable range $A$ is a mathematical abstraction of the possibility that Alice optionally reports the raw data. In Definition 2 an empty set $A$ corresponds to the particular case where all observables are intervals. A random set $A$ means individuals’ acceptable ranges vary (e.g., due to randomness by nature). On the other hand, a deterministic nonempty set $A$ means a fixed acceptable range uniformly for all individuals. From Subsection III-C and afterward, we will elaborate on the $A = \emptyset$ case and show that the population distribution is identifiable even without exact values of $Y$.

An alternative definition of privacy leakage is $L(A)$, meaning the probability of observing the exact value of $X$. Compared with the recommended $1 - \tau(M)$, the leakage here does not consider the intervals outside $A$. For example, in the particular case $A = \emptyset$, we have $L(A) = 0$, which is not appealing as the quantization also provides information.

### B. Fundamental Properties of Interval Mechanism

In this section, we show some desirable properties of canonical interval mechanisms. They can be directly extended to other mechanisms in later sections.

**Composition.** Suppose there are $k$ interval-private algorithms (or collectors), each querying the same data with a mechanism $M_j : Y \mapsto Z_j \in [1 : k]$. They may collaborate to narrow down the interval that contains a particular $Y$. This motivates the following ensemble mechanism, denoted by $\oplus_{j=1}^{k} M_j$, which is an interval mechanism induced by the intersections of anchor points and the union of acceptable ranges.

**Definition 3 (Ensemble Mechanism).** The ensemble of two privacy mechanisms

$\mathcal{M}_1 : X \mapsto Z = [Q_{[1]}, I(Q_{[1]}, Y), X \cdot 1_{X \in A_{[1]}}]$

with $j = 1, 2$ is defined by $\mathcal{M}_1 \oplus \mathcal{M}_2 :$

$X \mapsto Z = \{Q_{[1]} \oplus Q_{[2]}, I(Q_{[1]} \oplus Q_{[2]}, Y), X \cdot 1_{X \in A_{[1]} \cup A_{[2]}}\},$
where \( Q_1 \oplus Q_2 \) denotes the vector of all the anchor points from \( Q_1 \) and \( Q_2 \), and \( A_1 \cup A_2 \) denotes the union of two sets \( A_1, A_2 \). In general, the ensemble of \( k \) privacy mechanisms, denoted by \( \oplus_{i=1}^{k} M_i \), is recursively defined by \( \oplus_{i=1}^{k} M_i = (M_1 \oplus \cdots \oplus M_{k-1}) \oplus M_k \) \((k \geq 2)\).

**Theorem 4 (Composition Property).** Let \( M_1, \ldots, M_k \) be \( k \) interval mechanisms as in Definition 2. Then, we have
\[
1 - \tau(\oplus_{j=1}^{k} M_j) \leq \sum_{j=1}^{k} (1 - \tau(M_j)).
\]

An interpretation of the above theorem is that the privacy leakage of any ensemble mechanism is no larger than the sum of each of them. It is worth noting that \( Q^{[j]} \)'s may or may not be independent with each other, so communications between observers are allowed for this composition property to hold. In other words, this composition property holds even if the \( k \) mechanisms are adaptively chosen.

**Preprocessing.** Suppose that \( g : Y \mapsto g(Y) \) is a measurable function on \( Y \). Let \( A_g = \{g(y) : y \in A\} \) be the acceptable range for \( g(Y) \), which is carried over from \( A \). Suppose that an interval mechanism is applied to \( g(Y) \) instead of \( Y \) itself, with \( M : g(Y) \mapsto Z = [Q, I(Q, g(Y)), Y \cdot 1_{g(Y) \in A_g}] \). This corresponds to the ‘pullback’ privacy mechanism
\[
M_g : Y \mapsto Z_g = \{g^{-1}(Q), I(g^{-1}(Q), Y), Y \cdot 1_{Y \in A}\}.
\]
Here, \( g^{-1}(Q) \) denotes the partition of \( Y \) induced by the partition on \( g(Y) \) using \( Q \).

**Theorem 5 (Robustness to Preprocessing).** For any interval mechanism \( M \), it holds that \( \tau(M_g) \geq \tau(M) \), where the equality holds if and only if \( L(g^{-1}(A_g)) = L(A) \).

The above result shows that if a \( \tau \)-interval private observation is made on a transformation of \( Y \), namely \( g(Y) \), the privacy coverage of the raw data \( Y \) is not smaller than \( \tau \), or equivalently, the leakage at the raw data domain is no larger than \( 1 - \tau \). Furthermore, the \( M_g \) here may be regarded as \( M_j \) in Theorem 4, so Theorem 4 also holds for \( k \) observers that may target transformations of \( Y \) instead of \( Y \) itself. The inequality in Theorem 5 is strict when, e.g., \( Y \) is standard Gaussian, \( g(y) = y^2 \), and \( A = (-\infty, 0] \).

**Postprocessing.** The next result shows that the privacy leakage is not increased by subsequent processing of \( Z \).

**Theorem 6 (Robustness to Post-processing).** Suppose that \( M : Y \mapsto Z \) is an interval mechanism with \( \tau \)-interval privacy. Let \( f : Z \mapsto W \) be an arbitrary deterministic or random
mapping that defines a conditional distribution $W \mid Z$. Then $f \circ \mathcal{M} : Y \mapsto [Z, W]$ also meets $\tau$-interval privacy.

The above result is conceivable because $Y \rightarrow Z \rightarrow W$ is a Markov chain, and thus adding $W$ does not reveal more about the range of $Y$. We use $[Z, W]$ instead of $W$ in defining $f \circ \mathcal{M}$ because it is a complete observation. The amalgamation of composition property and robustness permits modular designs and analyses of interval mechanisms.

C. Extension: Interval Mechanism of General Topology

It is natural to extend the canonical interval mechanism in Subsection III-A by considering a partition of $\mathbb{R}$ into $m$ general ranges, denoted by $\{R^{(i)}\}_{i=1}^m$. We suppose each $R^{(i)}$ is a Borel set to properly define probability on them. Also, to operate data collection in practice, we let such a partition be determined by a fixed-dimension random vector $T \in \mathbb{R}^q$ and both $m, q$ be fixed positive integers. Formally, we introduce the following notion. We omit the acceptable range from now on for notational simplicity.

**Definition 4 (Extended Interval Mechanism).** Suppose that $T$ is a $q$-dimensional random vector independent with $Y$. Let

\[ R : t \mapsto \{R^{(i)}_t\}_{i=1}^m \]  

(5)

denote a map from each $t \in \mathbb{R}^q$ to a partition of $\mathbb{R}$. Let $I : (R, t, y) \mapsto i$ denote the indicator function defined by $y$ falling into $R^{(i)}_t$, $i \in [1 : m]$. A privacy mechanism, denoted by $\mathcal{M} : Y \mapsto Z$, maps $Y$ to $Z \triangleq [T, I(R, T, Y)]$.

A particular case is when $q = m - 1$, $T^{(i)} = Q^{(i)}$, and $R^{(i)}_T = (Q_{i-1}, Q_i]$ for $i \in [1 : q]$, which corresponds to Definition 2. It can be verified that the extended mechanism still satisfies the interval privacy in Definition 1 and all the properties in Subsection III-B. Next, we first explain why such extended mechanisms can be practically interesting. We then introduce the estimation of $F_Y$ and sufficient conditions to guarantee the distributional identifiability.

Consider a setting where we want to ensure a lower bound on each individual’s privacy coverage, namely $L(S_z) \geq \tau$ for each collected $z$ for a given $\tau > 0$. Note that a canonical mechanism cannot work without violating the distributional identifiability. To see that, let us
consider the left-most interval \((Q^{(0)}, Q^{(1)})\), where \(Q^{(0)}\) is fixed (e.g., \(-\infty\)) and \(Q^{(1)}\) is random. To ensure

\[ L((Q^{(0)}, Q^{(1)}) \geq \tau, \tag{6} \]

the smallest anchor point \(Q^{(1)}\) cannot take values in \((-\infty, F_{Y}^{-1}(\tau)]\). Then, the distribution of \(Y\) is not identifiable on the left tail. To address the issue, we consider the following example of Definition\(4\) that is not a canonical mechanism.

**Example 3.** Without loss of generality, suppose that \(T^{(1)} < \cdots < T^{(q)}\) almost surely. Let \(R_{T}^{(1)} = (-\infty, T^{(1)}] \cup (T^{(q)}, \infty)\), and \(R_{T}^{(i)} = (T^{(i-1)}, T^{(i)}]\) for \(i = 2, \ldots, q\). In other words, we concatenate the left-most and right-most canonical intervals into one range. This is naturally represented by a ‘ring’ topology as illustrated in Fig.\(7(a)\), where \(\pm\infty\) collapse into one anchor. In that figure, we used the notation of \(Q^{(i)}\) (instead of \(T^{(i)}\)) for an easier comparison with Fig.\(6(a)\). We also show a simple interface example in Fig.\(7(b)\), which is the counterpart of Fig.\(6(b)\). In this example, if \(Q\) is designed so that \(P_{Y}(Y \in (Q^{(1)}, Q^{(2)}]) \in [\tau, 1 - \tau]\) (which implicitly requires \(\tau \leq 0.5\)), it is intuitively possible to identify \(F_{Y}\). Next, we provide a formal method to guarantee the distributional identifiability. Our result is a nontrivial generalization of the existing theory for the Cases I&II interval data \[26\].

**Nonparametric maximum likelihood estimator (NPMLE):** We let \(B^{(i)}(T, Y) \overset{\Delta}{=} 1_{Y \in R_{T}^{(i)}}, \) or \(B^{(i)}\) for brevity, for \(i \in [1 : m]\). Let \(Z_{j} = [T_{j}, I(R, T_{j}, Y_{j})]\), or equivalently, \(Z_{j} = [T_{j}, B_{j}^{(1)}, \ldots, B_{j}^{(m)}]\), \(j \in [1 : n]\) denote the observed data, where \(n\) is the sample size. For any right-continuous distribution function \(F\) on \(Y\), let \(F(r)\) denotes the corresponding probability of a Borel set \(r\).
To estimate the underlying distribution of $Y$ without parametric assumptions\(^1\), we consider the log-likelihood functional
\[
\psi : F \mapsto \int_{\mathbb{R}^2 \times \mathcal{Y}} \left( \sum_{i=1}^{m} B^{(i)} \log F(R^{(i)}) \right) d\mathbb{P}_n(t, y)
\]
\[
= \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{m} B^{(i)}_j \log F(R^{(i)}_j),
\]  
(7)

where $\mathbb{P}_n(\cdot, \cdot)$ denotes the empirical probability measure from $[T_j, Y_j]$, $j \in [1 : n]$. We define NPMLE as a right-continuous distribution function $\hat{F}_n$ that maximizes $\psi(F)$. An example of the related quantities are visualized in Fig. 8. Note that the objective in (7) can only be defined up to the values of $\hat{F}_n$ at the anchor points from the intersections of observed ranges (e.g., $y_1, y_2, y_3$ in Fig. 8). As such, we consider the NPMLE as a piecewise function with only jumps at anchor points. Next, we provide conditions for interval mechanisms to preserve distributional information, namely distributional identifiability.

**Resolvability condition:** $T$ has a density with respect to the Lebesgue measure. For each $y$ in the closure of $\mathcal{Y}$, there is an open neighborhood $N(y)$ such that for all $y_a, y_b \in N(y)$, the interval $(y_a, y_b]$ satisfies: there exist $t, t'$ in the essential support of $T$ and ranges $r^{(k)} \in R_t, r^{(\ell)} \in R_{t'}$ ($1 \leq k, \ell \leq m$) such that $(y_a, y_b] \cap r^{(k)} = \emptyset$ and $(y_a, y_b] \cup r^{(k)} = r^{(\ell)}$.

The resolvability is determined by both the support of $T$ and topology $R$ introduced in (5). Intuitively speaking, a mechanism is resolvable if any small interval of $\mathcal{Y}$ can be a difference from two feasible ranges. It will be used in our proof in the following way. We will first prove that $\hat{F}_n(r) \approx F_Y(r)$ for each feasible range $r$. This, together with resolvability, gives $\hat{F}_n \approx F_Y$

\(^{1}\)If $F_Y$ is parameterized by a fixed-dimensional parameter, standard maximum likelihood estimation and asymptotics can be readily applied [27, Ch.5].
on any small interval \((y_a, y_b]\), which further implies \(\hat{F}_n \approx F_Y\) globally. For example, it can be verified that any canonical mechanism (Definition 2) is resolvable if its \(Q\) has a positive density wherever \(0 < F_Y(Q^{(1)}) < \cdots < F_Y(Q^{(m-1)}) < 1\). On the other hand, a canonical mechanism satisfying (6) is not resolvable, shown in Fig. 9(a). Also, a ring design in Example 3 is resolvable if \(Q^{(1)}\) has a positive density on \(\mathbb{R}\) and \(Q^{(2)}\) conditional on \(Q^{(1)}\) does not degenerate to a point, shown in Fig. 9(b).

We will also need the following condition. Let \(G \circ F\) denote the function that maps \(y\) to \(G(F(y))\). Recall the \(R, T, q\) in Definition 4. As before, with a slight abuse of notation, we use \(F(r)\) and \(F(y)\) to denote the probability of a Borel set \(r\) and the CDF at a point \(y\), respectively.

**Monotonicity condition:** For any CDF \(F\), the map \(R\) satisfies: a) for each \(j \in [1:q]\) and \(i \in [1:m]\), \(F(R_t^{(i)})\) is either non-decreasing or non-increasing in \(t^{(j)}\) with \(t^{(j')}(j' \neq j)\) fixed; b) there exist functions \(G^{(1)}, \ldots, G^{(q)}\) that are nondecreasing, continuous, and bounded on \([0,1]\) such that for each \(j \in [1:q]\), \(i \in [1:m]\), and \(t_a, t_b \in \mathbb{R}^q\) differing only in the \(j\)-th entry, we have \(|F(R_t^{(i)}) - F(R_t^{(i)})| \leq |G^{(j)}(F)(t^{(j)}) - G^{(j)}(F)(t^{(j)})|\).

Intuitively speaking, condition (a) states that the probability mass on each range is monotone in each entry of \(t\), and (b) means that the sensitivity of those probabilities can be controlled in terms of \(t\). It can be verified the above condition hold for all canonical mechanisms and the example in Fig. 7 with \(G\) being the identity map.

**Theorem 7.** Assume that an extended interval mechanism satisfies the above Resolvability and Monotonicity conditions, and \(F_Y\) is continuous. Then, \(\sup_{y \in Y} |\hat{F}_n(y) - F_Y(y)| \to 0\) almost surely as \(n \to \infty\).

**D. Extension: Individual-Level Privacy Enhancement**

In Subsection III-C, we considered the problem to ensure a lower bound on each individual’s privacy coverage. Specifically, for each individual \(z\) collected, we want to ensure that

\[ L(S_z) \geq \tau, \]  

where \(S_z\) denotes the corresponding range. Note that this requirement is for each individual, much stronger than lower-bounding population coverage. As shown in Example 3, a general approach is to design an extended interval mechanism such that the coverage of each feasible range is lower bounded, namely \(L(R_t^{(i)}) \geq \tau\) almost surely, \(\forall i \in [1:m]\). A limitation of this
approach is that it requires $\tau \leq 1/m$. What if the privacy system or an individual requires a large $\tau$, say 0.6? We propose an alternative approach below.

The key idea of the alternative approach is only to collect ranges that satisfy (8). Depending on practical needs, it can be implemented and interpreted in two ways.

**Way I**: We allow an individual to choose “Not wish to answer” as shown in Fig. 4(b), so $\tau$ is a representation of the (possibly unknown) underlying privacy sensitivity.

**Way II**: We let the system pick up the ranges satisfying (8), where $L(\cdot)$ can be approximated using an estimate of $F_Y$, and $\tau$ represents a known system-specific privacy budget.

Nevertheless, a skeptical individual may worry about information leakage from not collecting his/her data, especially when individuals are not de-identified (e.g., tracked by static IPs). This motivates the following interval mechanism. Let $\tau, \rho \in [0, 1]$ denote two constants, and $W \sim \text{Bern}(\rho)$ denote a Bernoulli random variable with $\Pr(W = 1) = \rho$.

**Definition 5 (Selective Mechanism).** With $Z$ in Definitions 1 or 4, a $(\tau, \rho)$-selective mechanism collects $Z$ if $L(S_Z) \geq \tau$ and $W = 1$ simultaneously hold, and ‘null’ otherwise, where $W \sim \text{Bern}(\rho)$ is independently generated.

Here, the term ‘null’ indicates that the system does not collect the range $Y$ belongs to, and it discards all the generated ranges for that individual. Also, the observed data $Z$ implicitly implies $L(S_Z) \geq \tau$. It can be verified that a selective mechanism satisfies interval privacy. We say that
a selective mechanism has an ‘ignorability’ property if the probability of $W = 0$ conditional on collecting ‘null’ is at least 0.5. Intuitively, not collecting data is likely due to an independently generated cutoff.

**Proposition 2.** A $(\tau, \rho)$-selective mechanism satisfies ignorability if $\rho \leq \mathbb{P}(L(S_Z) \geq \tau)$.

Intuitively, the smaller probability of meeting individual-level $\tau$-coverage, the smaller $\rho$ (less selection) needed for ignorability. Since $\mathbb{P}(L(S_Z) \geq \tau)$ is interpreted as the frequency of $L(S_Z) \geq \tau$, in line with the Way I or II, the collection system may estimate it using the response rate from historical collections, or calculate it from an estimated $F_Y$.

Regarding the distributional identifiability, Theorem 7 no longer applies because the collected ranges are selectively biased towards large coverages (at least $\tau$), causing dependence of the underlying data $Y$ and anchor $T$ (in Definition 4). We will show that one can still guarantee the distributional identifiability for selective mechanisms under an adaption of earlier results and an additional condition. Moreover, the above discussions assume a fixed $\tau$. To accommodate individuals’ heterogeneous privacy sensitivities, one may also consider a random $\tau$. Details on these are deferred to the Appendix.

Moreover, we previously considered one-time collection from each individual. In practice, when individuals have different but unknown privacy sensitivities, we consider the following extension of Way I. If an individual chooses “Not wish to answer,” the system adaptively proceeds with a further question to the same individual conditional on the previous one. A general mechanism was explained in Remark 1. We exemplify the idea below, also illustrated in Fig. 4(c). We will provide a real-data experiment in Subsection IV-E.

**Example 4.** Suppose that $Y \in [\underline{U}^{(0)}, \overline{U}^{(0)}]$. Let $G_{\underline{u}, \overline{u}}$ denote a distribution that is determined by $\underline{u}$ and $\overline{u}$, and supported on $[\underline{u}, \overline{u}]$. At each round $h \geq 1$, the system

1) generates $U^{(h)} \sim G_{\underline{U}^{(h-1)}, \overline{U}^{(h-1)}}$, $Z^{(h)} = [U^{(h)}, 1_{Y \leq U^{(h)}}]$;
2) lets $\overline{U}^{(h)} = U^{(h)}$ if $Y \leq U^{(h)}$, and $\underline{U}^{(h)} = U^{(h)}$ otherwise;
3) collects $\cup_{i \leq h} Z^{(h)}$ and proceeds to the next round only if $L(S_{Z^{(h)}}) \geq \tau$, and $\cup_{i \leq h} Z^{(h-1)}$ (‘null’ for $h = 1$) otherwise.
E. Regression with Private Responses

This section proposes a general approach to fit supervised regression using interval-private responses.

Suppose that we are interested in estimating the regression function with $Y$ being the response variable and $X \in \mathbb{R}^p$ the features. Suppose that $Y$ has been already privatized, and we only access an interval-private observation of $Y$ while the data $X$ is visible. This scenario occurs, for example, when Alice (who holds $Y$) sends her privatized data to Bob (who holds $X$) to seek Assisted Learning \cite{8}. The scenario also occurs when $Y$ has to be private while $X$ is already publicly available.

Following the typical setting of regression analysis, we postulate the data generating model $Y = f^*(X) + \varepsilon$, where $f^*$ is the underlying regression function to be estimated, and $\varepsilon \sim F_\varepsilon$ is an additive random noise. We suppose that $X$ is a random variable independent of $\varepsilon$, and that $F_\varepsilon$ is a known distribution, say Gaussian or Logistic distributions. We will discuss unknown $F_\varepsilon$ in Remark 5.

Suppose that a Case-I interval mechanism is used, and data are $n$ observations in the form of $D_i = [u_i, \delta_i, x_i]^T$, where $\delta_i \triangleq 1_{y_i \leq u_i}$, $i \in [1:n]$. (9)

Since $f^*$ is unknown, a general approach is to represent $f^*$ with linear functions $Y = X^T \beta + \varepsilon$, where $\beta \in \mathbb{R}^p$ is treated as an unknown parameter and $\varepsilon \sim F_\varepsilon$. The above model includes parametric regression and nonparametric regression based on series expansion (e.g., with polynomial, spline, or wavelet bases). To estimate $\beta$ from $D_i$’s, a classical way is to maximize the likelihood, e.g., $\beta \mapsto \prod_{i=1}^{n} [F_\varepsilon(u_i - x_i^T \beta)]^{\delta_i} \{1 - F_\varepsilon(u_i - x_i^T \beta)\}^{1-\delta_i}$ for Case-I intervals.

Though the likelihood approach is principled for estimating a parametric regression, its implementation depends on the specific parametric form of the regression function $f$, and its extension to nonparametric function classes is challenging. In supervised learning, data analysts typically use a nonparametric approach, such as various types of tree ensembles and neural networks. However, the existing regression techniques for point-valued responses ($Y$) cannot handle interval-valued responses. As such, we are motivated to ‘transform’ the interval-data format into the classical point-data form to enable direct uses of existing regression methods and software.
Our main idea is to transform the data format from intervals to point values so that many existing regression methods can be readily applied. We propose to use

\[ \tilde{Y} = \mathbb{E}(Y \mid D, X) \]  

(10)
as a surrogate to \( Y \). This is motivated from the observation that \( \tilde{Y} \) is an unbiased estimator of \( f(x) \) for a given \( X = x \), namely \( \mathbb{E}(\tilde{Y} \mid x) = \mathbb{E}(Y \mid x) = f(x) \). Suppose that we choose a loss function \( \ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R} \) such as \( L_1 \) or \( L_2 \) loss. For notational convenience, we let \( \mathbb{E}_n \) denote the empirical expectation, e.g., \( \mathbb{E}_n \ell(\tilde{Y}, f(X)) = n^{-1} \sum_{i=1}^{n} \ell(\tilde{Y}_i, f(X_i)) \). Based on the above arguments in (10), it is desirable to solve the following optimization problem

\[
\min_{f \in \mathcal{F}} \mathbb{E}_n \ell(\tilde{Y}, f(X)) 
\]

(11)

where \( \tilde{Y} = \mathbb{E}(Y \mid D, X) = f(X) + \mathbb{E}(\varepsilon \mid D, X) \).

The following result justifies the validity of using \( \tilde{Y} \) as a surrogate of \( Y \) to estimate \( f \), if the \( \tilde{Y} \) in (12) were statistics. We define the norm of \( f \) to be \( \|f\|_{P_X} \triangleq \mathbb{E}(f(X)^2) \). Suppose that the underlying regression function \( f^* \) belongs to \( \mathcal{F} \), a parametric or nonparametric function class with bounded \( L_2(P_X) \)-norms. The following result shows that the optimal \( f \) obtained by minimizing (11) with a squared loss \( \ell \) is asymptotically close to the underlying truth \( f^* \). Extensions of the proof to other loss functions are left as future work.

**Theorem 8 (Regression Estimation).** Suppose that

\[
\sup_{f \in \mathcal{F}} \left| \mathbb{E}_n(\tilde{Y} - f(X))^2 - \mathbb{E}(\tilde{Y} - f(X))^2 \right| \to_p 0
\]

(convergence in probability) as \( n \to \infty \). Then, any sequence \( \hat{f}_n \) that maximizes \( \mathbb{E}_n(\tilde{Y} - f(X))^2 \) converges in probability to \( f^* \) in the sense that \( \|\hat{f}_n - f^*\|_{P_X} \to_p 0 \) as \( n \to \infty \).

In practice, however, the calculation of \( \tilde{Y} \) itself is unrealistic as it involves the knowledge of \( f(X) \). In other words, the unknown function \( f \) appears in both the optimization (11) and calculation of surrogates (12). The above difficulty motivates us to propose an iterative method where we iterate the steps in (11) and (12), using any commonly used supervised learning method to obtain \( \hat{f} \) at each step.

The pseudocode is provided in Algo. 1, where Case-I intervals are considered for brevity. In practice, we set the initialization by \( \hat{f}_{n,0}(x) = 0 \) for all \( x \). We experimentally found that Algo. 1
Algorithm 1 Interval Regression by Iterative Transformations (a Case-I example)

**Input:** Interval-valued responses $D_i$ in Eq. (9) and predictors $x_i \in \mathbb{R}^p$, $i \in [1:n]$, function class $F$, error distribution function $F_{\varepsilon}$

**Initialization:** Round $k = 0$, function $\hat{f}_{n,0}(\cdot)$

1: repeat
2: Let $k \leftarrow k + 1$
3: Update the representative $\tilde{y}_i, i \in [1:n]$
   $$\tilde{y}_i = \hat{f}_{n,k-1}(x_i) + \mathbb{E}(\varepsilon | \tilde{u}_i, \delta_i, x_i)$$  \hspace{1cm} (13)
   where $\tilde{u}_i = u_i - \hat{f}_{n,k-1}(x_i)$.
4: Fits a supervised model $\hat{f}_{n,k}$ using $(\tilde{y}_i, x_i)$ as labeled data by optimizing (11) using a preferred method
5: until A stop criterion satisfied (e.g., if the fitted values do not vary much)

**Output:** The estimated function $\hat{f}_{n,k} : \mathbb{R}^p \rightarrow \mathbb{R}$

is robust and works well for a variety of nonlinear models such as tree ensembles and neural networks.

**Remark 5 (Computing the conditional expectation in (13)).** The term $\mathbb{E}(\varepsilon | \tilde{u}_i, \delta_i, x_i)$ is essentially $\mathbb{E}(\varepsilon | \varepsilon \leq u_i - \hat{f}_{k-1}(x_i))$ if $\delta_i = 1$, or $\mathbb{E}(\varepsilon | \varepsilon > u_i - \hat{f}_{k-1}(x_i))$ otherwise. When calculating (13), we need to specify a distribution for the error term $\varepsilon$. Though a misspecified distributional assumption often affects inference results [28], we found from experimental studies that the accuracy of estimating $f$ here is not sensitive to misspecification of the noise distribution (see Subsection IV-D). A practical suggestion to data analysts is to treat $\varepsilon$ as Logistic random variables to simplify the computation. Also, if the standard deviation of the noise $\varepsilon$ is unknown in practice, we suggest estimate $\sigma^2$ with $n^{-1} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$ at each iteration of Algo. 1

IV. Experiments

We provide experiments on the use of interval privacy, including unsupervised, supervised, and real-data examples.

A. Estimation of Moments

In this experiment, we demonstrate moment estimation with interval-private data of reasonably broad privacy coverage. Suppose that 100 data are generated from $Y \sim \mathcal{N}(0.5, 1)$, and the private data are based on the Case-I mechanism with $U \in \text{Uniform}[-T, T]$, where $T = 2n^{1/3}$. The privacy coverage is around 0.95 (or 5% leakage). The goal is to estimate $\mathbb{E}(Y)$. We consider two
|                      | Estimate $\mathbb{E}(Y)$ | Estimate $\mathbb{E}(Y^2)$ |
|----------------------|--------------------------|---------------------------|
|                      | $n$ 0% 1% 5% 0% 1% 5%     |                           |
| Private data         | 100 0.45 0.44 0.58 0.79 0.82 1.03 |
| (Example 2)          | 1000 0.29 0.33 0.99 0.57 0.64 1.96 |
| Private data         | 100 0.32 0.36 0.92 13.09 11.58 10.64 |
| (NPMLE)              | 1000 0.12 0.21 1.13 3.68 4.08 4.45 |
| Raw data             | 100 0.08 9.98 49.93 0.14 9.98 49.89 |
| (Mean)               | 1000 0.03 9.98 49.92 0.04 9.98 49.88 |
| Raw data             | 100 0.11 0.10 0.12 0.66 0.65 0.60 |
| (Median)             | 1000 0.03 0.03 0.07 0.67 0.65 0.60 |

methods and compare them with the baseline estimates using raw data (in hindsight) in Table II. The first method, denoted by ‘Example 2’, uses the estimator in (3). By a similar argument as the proof of Proposition I, the choice of $T$ guarantees that the $\mu$ can be consistently estimated. The second method is the NPMLE implemented in the ‘Icens’ R package [24]. We also consider two methods based on raw data: the sample average and the sample median. To demonstrate the robustness, we add 0%, 1%, and 5% proportion of outliers (meaning $Y = 999$). We also consider the estimation of $\mathbb{E}(Y^2)$ in a similar setting, except that we use $U \in \text{Uniform}[0, 2T]$ to collect $Y^2$ so that the privacy coverage remains around 0.95.

The results summarized in Table II indicate that the estimation under highly private data is reasonably well when compared with the oracle approach with 0% outlier. Also, the estimation from interval-private data tends to be more robust against outliers than the estimation based on the simple mean and comparable to the median (using raw data).

**B. Estimation of Regression Functions**

We first demonstrate the method proposed in Subsection III-E on the linear regression model $Y = f(X) + \varepsilon$, where $f(X) = \beta X$ is to be estimated from the Case-I privatized data $Z = [U, \Delta]$. 

We generate \( n = 200 \) data with \( \beta = 1, X, \varepsilon \sim \text{i.i.d. } \mathcal{N}(0,1), U \) a Logistic random variables with scale 5. The corresponding privacy coverage is around 0.9. Fig. 10 demonstrate a typical result. The prediction error is evaluated by mean squared errors \( E(f(\tilde{X}) - \hat{f}(\tilde{X}))^2 \) where \( \tilde{X} \) denotes the unobserved (future) data. With a limited size of data, the algorithm will produce an estimate \( \hat{f}(x) = \hat{\beta}X \) that converges well within 20 iterations. The initialization is done by simply setting \( \hat{f}_{n,0}(x) = 0 \).

In another experiment, we demonstrate the method proposed in Subsection III-E on the nonparametric regression model \( Y = f(X) + \varepsilon \), using the Case-II privatized data \( Z = [U, V, \Delta, \Gamma] \).

Suppose that \( n = 200 \) data are generated from quadratic regression \( f(X) = X^2 - 2X + 3 \), and \( \varepsilon \sim \mathcal{N}(0,1) \). Let \( U = \min(L_1, L_2), V = \max(L_1, L_2) \), where \( L_1, L_2 \) are independent Logistic random variables with scale 5. The corresponding privacy coverage is around 0.9. Random Forest (depth 3, 100 trees) with features \( X_1 = X, X_2 = X^2 \) are used to fit Algo. 1. Fig. 11 demonstrates a typical result. With a limited data size, the algorithm can produce a tree ensemble that converges well within 20 iterations.

C. Tradeoff between Learning and Privacy Coverage

The tradeoff between privacy coverage and learning performance is computable often in parametric settings, where the asymptotic variance and coverage privacy can be treated as functions of distribution parameters, and in some nonparametric learning contexts (see, e.g., Theorem 2 and relevant discussions). In an experiment, we demonstrate the tradeoff with \( n = 200 \) data as used in the first experiment of Subsection IV-B. We consider the Case-I mechanism, where \( U \) is generated from Logistic distributions with scales 0.1, 0.3, 0.5, 1, 3, 5, 10, 20, and 30. We numerically compute the prediction errors and privacy coverages. The results, summarized in Fig. 12 indicate that the performance is not sensitive to privacy coverage unless the latter is very close to one.

D. Sensitivity of Misspecified Noise

We empirically found that the estimation accuracy is generally not much affected by a misspecified distribution of \( \varepsilon \) when calculating (13). We demonstrate the sensitivity of wrongly specifying a distribution term using a specific example. A more sophisticated sensitivity analysis is left as future work. We generate data in the same way as in Subsection IV-B except that the actual noise follows \( t \)-distributions with degrees of freedom \( d = 1000, 100, 10, 5, 3, \) and 1. Here,
Fig. 10. Experiments in Subsection [V-B] Snapshots of Algo. [I] for linear regression at the 1st (left-top), 3rd (right-top), and 20th (left-bottom) iterations, and the prediction error ($L_2$ loss) versus iteration (right-bottom). Grey vertical segments indicate the observed intervals in the form of $(−\infty, u]$ or $(u, \infty)$; Blue dots and lines indicate the unprotected data $Y$ and the underlying true regression function; Red dots and dashed lines indicate the adjusted data $\tilde{Y}$ in (10) and the estimated regression function.

$d = 1000$ is virtually Gaussian while $d = 1$ corresponds to a (heavy-tailed) Cauchy distribution. The postulation is still a Gaussian noise (so that it is misspecified). The results summarized in Fig. [13] indicates that the performance (evaluated by the mean squared error) is not severely affected, and less deviation tends to produce less degradation in performance.
E. Case Study: Distribution from Individual-Adaptive Surveys

We developed a web-based survey system and deployed it on MTurk to collect interval-private data. We de-identified the voluntary participants and randomized them into two groups. In the first group, each anonymous participant was asked privacy-sensitive questions in the form of Fig. 4(c), where the progressive mechanism was based on Example 4. In particular, $[U^{(0)}, \bar{U}^{(0)}]$ was specified according to the question and each $U^{(h)}$ was uniformly generated.
Fig. 12. Experiments in Subsection IV-C: The prediction error versus privacy coverage (left), and privacy coverage versus the spread of intervals, as measured by the standard deviation of $U$ (right). The shaded bands indicate ±standard errors from 50 replications.

We use the ultimate intervals from the progressive mechanism (“Round-X”) to obtain the NPMLE $\hat{F}_n$. Using the empirical distribution of the point data collected from the second group to approximate the population $F_Y$, we calculate the squared Energy Distance $\text{ED}(\hat{F}_n, F_Y) \triangleq \int_R (\hat{F}_n(y) - F_Y(y))^2dy$ as the estimation error, and record the privacy coverage. We repeat the above to only the data collected from the first round of questions (“Round-1”). As we can see from Tab. III, the progressive mechanism tends to reduce estimation error by adapting to individual-level privacy sensitivities.
Fig. 13. Performance (mean squared error) versus misspecification level (in terms of the $t$-degree of freedom $d$). A larger $d$ means less misspecification. The bands indicate ±standard errors from 200 independent replications.

### TABLE III

**Subsection IV-E** EXPERIMENT: DISTRIBUTION ESTIMATION AND COVERAGE FROM NON-PROGRESSIVE ("Round-1") AND PROGRESSIVE INTERVAL MECHANISMS ("Round-X") FOR THREE QUESTIONS.

| $Y$          | ED($F_n, F$) | Coverage |
|--------------|--------------|----------|
|              | Round-1      | Round-X  | Round-1 | Round-X |
| Salary       | [0, 150] ($\times$1k) | 1.38 | 0.71 | 0.68 | 0.35 |
| Cash         | [0, 300] ($\times$1k) | 1.72 | 0.87 | 0.75 | 0.45 |
| Intercourse  | [0, 100] ($\times$1) | 0.73 | 0.40 | 0.89 | 0.63 |

**F. Case Study: Life Expectancy Regression**

In the experimental study, we considered the ‘life expectancy’ data from the kaggle open-source dataset [18], originally collected from the World Health Organization (WHO). The data consist of 193 countries from 2000 to 2015, with 2938 data items/rows uniquely identified by the country-year pair. The learning goal is to predict life expectancy using 20 potential factors,
such as demographic variables, immunization factors, and mortality rates.

We will exemplify the use of Algo. 1 under three mechanisms. The first mechanism (‘Oracle’) uses the raw data of \( Y \) (life expectancy). The second mechanism (‘\( \mathcal{M}_1 \)’) is described by \( Y \mapsto Z \), where \( Z \) is in the form of \( [U - 1, U + 1] \), \( U \) is generated from the Logistic distribution with scale 1, and \( A = \{ y : U - 1 < y \leq U + 1 \} \). An interpretation is that during the data collection, individual data with an overly short or long life expectancy tend to be reported as half-interval (namely \( \leq U - 1 \) or \( > U - 1 \)), while those within the mid-range \( A \) tend to be exactly reported. The third mechanism (‘\( \mathcal{M}_2 \)’) is a Case-II mechanism described by (4), where \( Q = [U, V] \) is generated from the ordered Logistic distribution with scale 2. The last mechanism (‘\( \mathcal{M}_3 \)’) is a Case-I mechanism where \( Q = U \) is generated from the Logistic distribution with scale 5. The interpretation of \( \mathcal{M}_2 \) or \( \mathcal{M}_3 \) is that individual data are quantized into random categories. We calculate the privacy coverage (Definition 1) for each privacy mechanism using the empirical distribution and summarize it in Table IV.

For each mechanism, the predictive performance of the fitted regression under three methods, namely linear regression (LR), gradient boosting (GB), and random forest (RF), are evaluated using the five-fold cross-validation. The performance results are summarized in Table IV. The results show that a privacy mechanism with smaller privacy coverage tends to perform better, which is the expected phenomenon due to privacy-utility tradeoffs. The results also show a (statistically) negligible performance gap between \( \mathcal{M}_1 \), \( \mathcal{M}_2 \), and the Oracle (meaning that the raw data are used). The performance starts to degenerate only in the last mechanism, where there is a large privacy coverage (94%) or small privacy leakage (6%).

To visualize the data and individual-level privacy coverage, we also show a snapshot of the database in Tab. I. There, we used the Case-II mechanism and generated \( U, V \) from the standard Logistic distribution.

V. RELATED LITERATURE

This section reviews some other perspectives of data privacy.

Database privacy. A popular framework of evaluating data privacy is through differential privacy [29], [30], a cryptographically motivated definition of privacy to protect the existence of an individual identity in a database [31]–[35]. A database is a matrix whose rows represent individuals and columns represent their attributes. Differential privacy measures privacy leakage by a parameter \( \varepsilon \) that bounds the likelihood ratio of the output of an algorithm under two
### TABLE IV

**Subsection IV-F** Experiment: Predictive performance of linear regression (LR), gradient boosting (GB), and random forest (RF) methods under different mechanisms, evaluated by the $R^2$ and mean absolute error (MAE) from 5-fold cross validations.

| Coverage | Oracle | $\mathcal{M}_1$ | $\mathcal{M}_2$ | $\mathcal{M}_3$ |
|----------|--------|-----------------|-----------------|-----------------|
| LR       | $R^2$  | 0.79(0.02)      | 0.78(0.02)      | 0.78(0.02)      | 0.52(0.09)      |
|          | MAE    | 3.21(0.07)      | 3.25(0.08)      | 3.23(0.04)      | 4.43(0.25)      |
| GB       | $R^2$  | 0.89(0.01)      | 0.86(0.01)      | 0.85(0.01)      | 0.74(0.01)      |
|          | MAE    | 2.3(0.18)       | 2.56(0.11)      | 2.68(0.09)      | 3.65(0.13)      |
| RF       | $R^2$  | 0.82(0.02)      | 0.78(0.02)      | 0.75(0.02)      | 0.69(0.02)      |
|          | MAE    | 2.86(0.12)      | 3.25(0.09)      | 3.41(0.14)      | 3.97(0.05)      |

databases differing in a single individual. The standard tool for creating differential privacy is the sensitivity method \[29\], which first computes the desired algorithm output (e.g., the count) from the database, and then adds noise proportional to the largest possible change induced by modifying a single row in the database. Differential identifiability \[36\] was developed as an alternative formulation to guarantee differential privacy, based on the probability of individual identification conditional on the output. This notion was also extended to the identifiability of databases instead of individuals \[37\].

**Sanitization.** In some applications such as census publication, the output of operating on a database needs to be an anonymized and perturbed version of the original database (also known as ‘sanitization’) to protect individual privacy. Along this direction, a classical approach is based on the notion of $k$-anonymity \[38\], meaning that for every individual, there exist $k-1$ others with the same tuple of non-private attribute values (assumed to exist) for a pre-specified $k$. Since $k$-anonymity does not necessarily protect private attributes, there have been some extensions such as the $t$-closeness \[39\]. Moreover, information-theoretic quantities such as mutual information and average distortion have been used to quantify privacy in database sanitization (see, e.g., \[37\], \[40\]–\[43\]).

**Local privacy.** The main difference between database privacy and local privacy (focus of this work) is summarized below. First, local privacy protects each data value or the associated
individual identity during data collection, while database privacy protects the presence of an individual in an already-collected database. Second, local privacy is supposed to disclose or collect individual-level data, while database privacy is developed for querying summary statistics. Third, in practical implementations, database privacy involves three parties: data owners (individuals), a data collector (trusted third party, often an organization) who maintains the database, and analysts who query statistics from the database. Local privacy may only involve two parties: data owners and an (untrusted) data collector who may immediately analyze the collected data. Compared with database privacy, local privacy is much less studied in the literature. Interval privacy can be regarded as a framework for local privacy.

*Local differential privacy.* An existing notion of local privacy is local differential privacy [14], [15], [44], [45], a local version of differential privacy [29] where only randomized data are available to data collectors and analysts. Local differential privacy is a criterion that restricts the conditional distributions of the privatized data on any two different raw data to have a density ratio close to one, often realized by perturbing the raw data with additive noise. We exemplify the difference between interval privacy and local differential privacy through a toy example. Consider a salary of $25k and another salary of $250k. In a typical differential privacy scenario, the two private values are obfuscated with two random point values whose distributions exhibit epsilon-difference in density ratios. In contrast, under an interval privacy mechanism, the two private values are obfuscated with two random intervals, say (0, $100k) and [$200k, $\infty$). An operational difference is that interval privacy offers information by narrowing down the support size, while local differential privacy offers information by perturbing the value. A conceptual difference is that interval privacy ensures an adversary does not gain additional information of $Y$ on a large support based on its collected data $Z$ and prior knowledge on $Y$ (through posterior ratio of $Y \mid Z$), while local differential privacy limits the additionally gained information through the likelihood ratio of $Z \mid Y$.

**VI. Conclusion and Further Remarks**

In this work, we developed concepts, properties, and some use scenarios of interval privacy. We demonstrated the interval mechanism through various specific designs. Future work includes the following directions. First, it is worth extending the notion of interval privacy for continuously-valued variables to discrete variables, e.g., categorical, ordinal, and count data, and from one-dimensional to multi-dimensional ranges. Second, it would be interesting to apply the developed
concepts and methods to tackle many supervised, unsupervised, and reinforcement learning tasks. Third, a more sophisticated analysis of the tradeoff between privacy and learning utility in particular mechanisms deserves further study.

The Appendix contains further technical details.

REFERENCES

[1] P. Voigt and A. Von dem Bussche, “The EU general data protection regulation (GDPR),” A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017.
[2] N. Evans, S. Marcel, A. Ross, and A. B. J. Teoh, “Biometrics security and privacy protection,” IEEE Signal Process. Mag., vol. 32, no. 5, pp. 17–18, 2015.
[3] M. S. Cross and A. Cavallaro, “Privacy as a feature for body-worn cameras,” IEEE Signal Process. Mag., vol. 37, no. 4, pp. 145–148, 2020.
[4] Google, “Google security whitepaper,” shorturl.at/lzMT4 Jan 2019.
[5] Facebook, “Communicating about privacy: Towards people-centered and accountable design,” shorturl.at/hpIJ8 July 2020.
[6] M. M. Alabbadi, “Mobile learning (mlearning) based on cloud computing: mlearning as a service,” in UbiComp, 2011.
[7] M. Ribeiro, K. Grolinger, and M. A. Capretz, “Mlaas: Machine learning as a service,” in ICMLA. IEEE, 2015, pp. 896–902.
[8] X. Xian, X. Wang, J. Ding, and R. Ghanadan, “Assisted learning: A framework for multi-organization learning,” Proc. NeurIPS, 2020.
[9] J. Zhou, J. Ding, K. M. Tan, and V. Tarokh, “Model linkage selection for cooperative learning,” J. Mach. Learn. Res., 2021.
[10] M. P. Couper, M. W. Traugott, and M. J. Lamias, “Web survey design and administration,” Public Opin. Q., vol. 65, no. 2, pp. 230–253, 2001.
[11] M. S. Litwin and A. Fink, How to assess and interpret survey psychometrics. Sage, 2003, vol. 8.
[12] H. Wang, E. Skau, H. Krim, and G. Cervone, “Fusing heterogeneous data: A case for remote sensing and social media,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 12, pp. 6956–6968, 2018.
[13] M. Sun and W. P. Tay, “On the relationship between inference and data privacy in decentralized iot networks,” IEEE Trans. Inf. Forensics Secur., vol. 15, pp. 852–866, 2019.
[14] A. Evfimievski, J. Gehrke, and R. Srikant, “Limiting privacy breaches in privacy preserving data mining,” in Proc. SIGMOD/Pods03, 2003, pp. 211–222.
[15] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, “What can we learn privately?” SIAM J. Comput., vol. 40, no. 3, pp. 793–826, 2011.
[16] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, “Minimax optimal procedures for locally private estimation,” J. Am. Stat. Assoc., vol. 113, no. 521, pp. 182–201, 2018.
[17] M. Sun, W. P. Tay, and X. He, “Towards information privacy for the internet of things,” arXiv preprint arXiv:1611.04254, 2016.
[18] Kaggle, “Life expectancy dataset,”https://tinyurl.com/yxgaa4go, 2020.
[19] S. K. Orr, “New technology and research: An analysis of internet survey methodology in political science,” PS Polit. Sci. Polit., vol. 38, no. 2, pp. 263–267, 2005.
[20] Z. Zhang and J. Sun, “Interval censoring,” Stat. Methods Med. Res., vol. 19, no. 1, pp. 53–70, 2010.
[21] C. Gentry, “Fully homomorphic encryption using ideal lattices,” in Proc. STOC, 2009, pp. 169–178.
[22] P. Groeneboom and J. A. Wellner, Information bounds and nonparametric maximum likelihood estimation. Springer Science & Business Media, 1992, vol. 19.
[23] B. W. Turnbull, “The empirical distribution function with arbitrarily grouped, censored and truncated data,” J. Royal Stat. Soc. B, vol. 38, no. 3, pp. 290–295, 1976.
[24] R. Gentleman and A. Vandal, “Icens: Npmle for censored and truncated data,” R package version, vol. 1, no. 0, 2010.
[25] R. Geskus and P. Groeneboom, “Asymptotically optimal estimation of smooth functionals for interval censoring, case 2,” Ann. Stat., vol. 27, no. 2, pp. 627–674, 1999.
[26] P. Groeneboom, “Nonparametric maximum likelihood estimators for interval censoring and deconvolution,” 1991.
[27] A. W. Van der Vaart, Asymptotic statistics. Cambridge university press, 2000, vol. 3.
[28] J. Ding, V. Tarokh, and Y. Yang, “Model selection techniques: An overview,” IEEE Signal Process. Mag., vol. 35, no. 6, pp. 16–34, 2018.
[29] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in Theory of cryptography conference. Springer, 2006, pp. 265–284.
[30] C. Dwork, “Differential privacy,” Encyclopedia of Cryptography and Security, pp. 338–340, 2011.
[31] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, “Differentially private empirical risk minimization,” J. Mach. Learn. Res., vol. 12, no. 3, 2011.
[32] A. D. Sarwate and K. Chaudhuri, “Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data,” IEEE Signal Process. Mag., vol. 30, no. 5, pp. 86–94, 2013.
[33] J. Dong, A. Roth, and W. J. Su, “Gaussian differential privacy,” arXiv preprint arXiv:1905.02383, 2019.
[34] M. Neunhoeffer, Z. S. Wu, and C. Dwork, “Private post-gan boosting,” arXiv preprint arXiv:2007.11934, 2020.
[35] G. Vietri, G. Tian, M. Bun, T. Steinke, and Z. S. Wu, “New oracle-efficient algorithms for private synthetic data release,” arXiv preprint arXiv:2007.05453, 2020.
[36] J. Lee and C. Clifton, “Differential identifiability,” in Proc. KDD, 2012, pp. 1041–1049.
[37] W. Wang, L. Ying, and J. Zhang, “On the relation between identifiability, differential privacy, and mutual-information privacy,” IEEE Trans. Inf. Theory, vol. 62, no. 9, pp. 5018–5029, 2016.
[38] L. Sweeney, “k-anonymity: A model for protecting privacy,” Int. J. Uncertain. Fuzziness Knowl. Syst., vol. 10, no. 05, pp. 557–570, 2002.
[39] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” in Proc. ICDE. IEEE, 2007, pp. 106–115.
[40] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, “From t-closeness-like privacy to postrandomization via information theory,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 11, pp. 1623–1636, 2009.
[41] F. du Pin Calmon and N. Fawaz, “Privacy against statistical inference,” in Proc. Allerton, 2012, pp. 1401–1408.
[42] L. Sankar, S. R. Rajagopalan, and H. V. Poor, “Utility-privacy tradeoffs in databases: An information-theoretic approach,” IEEE Trans. Inf. Forensics Secur., vol. 8, no. 6, pp. 838–852, 2013.
[43] A. Makhdoumi and N. Fawaz, “Privacy-utility tradeoff under statistical uncertainty,” in Proc. Allerton. IEEE, 2013, pp. 1627–1634.
[44] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, “Local privacy and statistical minimax rates,” in IEEE Symp. FOCS, 2013, pp. 429–438.
[45] A. D. Sarwate and L. Sankar, “A rate-distortion perspective on local differential privacy,” in Proc. Allerton. IEEE, 2014, pp. 903–908.