Information Sciences 626 (2023) 537–558

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Total variation distance privacy: Accurately measuring inference attacks and improving utility

Jingyu Jia [a,d], Chang Tan [a,d], Zhewei Liu [b,d], Xinhao Li [b,d], Zheli Liu [b,d], Siyi Lv [b,d,*], Changyu Dong [c]

[a] College of Computer Science, Nankai University, Tianjin, China
[b] College of Cyber Science, Nankai University, Tianjin, China
[c] Institute of AI and Blockchain, Guangzhou University, Guangzhou, China
[d] Tianjin Key Laboratory of Network and Data Security Technology, Nankai University, China

ARTICLE INFO

Article history:
Received 24 October 2022
Received in revised form 7 December 2022
Accepted 2 January 2023
Available online 7 January 2023

Keywords:
Differential privacy
Total variation distance
Membership inference attacks

ABSTRACT

Differential privacy (DP) is a general approach to defend against inference attacks, but hard to balance the privacy-utility trade-off for some complex data analysis tasks. To improve the utility of data analysis, a weaker privacy definition with a more accurate estimate of privacy risk may be a straightforward and effective solution. Total variation distance (TVD) privacy is an appropriate tool for this issue, but it has not been adequately studied. In this paper, we systematically study TVD privacy and explore its applications. We provide a comprehensive theoretical analysis of TVD privacy and demonstrate its advantage in measuring privacy risks with the example of membership inference attacks. Our work indicates that TVD privacy is a helpful tool in estimating privacy risks and has the potential to be widely used as a general privacy definition.

1. Introduction

Inference attack, which aims to infer private information from data analysis tasks, is a significant threat to privacy-preserving data analysis. Many kinds of inference attacks have been proposed, such as Membership Inference Attacks (MIAs) [1,2], attribute inference attacks [3,4], and category inference attacks [5]. For example, MIAs train an inference model to distinguish the difference in predictions of the target model for trained and non-trained inputs, determining whether a given record is a training sample. These attacks have been shown to cause huge privacy risks to services. Shokri et al. [1] have successfully implemented MIAs for training services provided by Google and Amazon.

Differential privacy (DP) [6–13], the gold standard for privacy protection, is considered an effective approach to defend against inference attacks. It requires that the privacy algorithms ensure that changes in individual records have little impact on the analysis results. Since DP guarantees that analyzers have difficulty distinguishing whether a given record is in datasets. DP mechanisms are particularly well suited for defending against MIAs, intuitively. Many works [8,10,11,13] investigated the effectiveness of DP mechanisms in defending against MIAs. Their works, both the theoretical upper bound on privacy loss provided by DP and the empirical lower bound on privacy leakage of inference attacks, offer a theoretical basis for designing and using privacy mechanisms in the real world.

* Corresponding author.
E-mail address: lvsiyi@nankai.edu.cn (S. Lv).

https://doi.org/10.1016/j.ins.2023.01.037
0020-0255/

DP provides formal privacy guarantees but sometimes leads to unacceptable accuracy loss in data analysis. Recently, some studies had negative views about DP machine learning mechanisms. Nasr et al. [8] verified that the upper bound of privacy leakage for DP is tight when the adversary in machine learning has sufficient capabilities. In other words, for the existing DP mechanisms in machine learning, it is difficult to reduce the noise by improving the theoretical analysis. Jayaraman et al. [11] found that DP mechanisms have difficulty providing meaningful privacy guarantees while providing limited accuracy loss in machine learning. These works reveal that it is difficult for the DP mechanisms to satisfy both the privacy and utility needs of the model, and the issue is difficult to solve through theoretical analysis.

The above negative findings can be attributed to the strict privacy standards of DP. It is sometimes difficult to avoid large noise in data analysis tasks for personal privacy. In addition, DP may overestimate some privacy risks. DP calculates a theoretical upper bound on the privacy risk based on the worst-case event of the mechanism distribution. However, the worst-case event has a low probability for most distributions, such as the Gaussian distribution. DP can provide tight upper bounds for some indicators that focus on worst-case events. However, for other indicators, such as the accuracy of MIAs, DP may give loose measures and add excessive noise.

A weaker privacy definition with a more accurate estimate of privacy risks may be a straightforward and effective solution to improve the utility of data analysis. Total variation distance (TVD) privacy is an appropriate definition for this issue. It was proposed as a weaker definition of privacy than DP, with more accurate estimates for some privacy risks. Recently, some work [14,15] began demonstrating TVD's potential application in measuring privacy risks. However, these works selectively analyzed or exploited partial properties of TVD, while a comprehensive study of TVD privacy has not emerged. In this paper, we systematically study TVD privacy and explore its applications.

Our work:

Comprehensive analysis of TVD privacy: We provide a comprehensive theoretical analysis of TVD privacy. We propose the protection target of TVD privacy and provide several formal theorems, such as the post-processing theorem, composition theorem, and privacy amplification theorem. These studies provide the theoretical basis for TVD privacy and demonstrate that TVD privacy can be applied as a general privacy definition in real scenarios.

Private mechanisms analysis: We extend the application of TVD privacy. Specifically, we use TVD privacy to analyze the Laplace and Gaussian mechanisms in single-dimensional numerical queries and the Gaussian mechanisms in multi-dimensional numerical queries. TVD privacy can provide precise measures of these private mechanisms. This work contributes to using TVD privacy as a general definition in many fields.

Privacy risk estimation: We demonstrate the advantage of TVD privacy in measuring privacy risks with the example of MIAs. We first use TVD privacy to analyze the DP machine learning mechanism [6]. Then, we analyze four indicators (accuracy of MIAs, membership advantage, ROC curve, and positive predictive value) commonly used in MIAs. We prove that TVD privacy can accurately estimate the accuracy of MIAs, thereby improving the utility of the training model. Under the same MIA accuracy constraints, TVD privacy can reduce noise by more than 50% compared to DP. For the other metrics, TVD privacy and DP estimates are complementary. By combining TVD privacy and DP, users can more accurately understand the privacy risks of machine learning. We provide source code [1] for computing TVD privacy to help users set appropriate parameters in private machine learning.

In a word, our work demonstrates that TVD privacy is a helpful tool in estimating privacy risks and has the potential to be widely used as a general privacy definition.

2. Preliminary

2.1. Differential Privacy

Differential privacy (DP) is a privacy definition for privacy-preserving data analysis. It constrains the impact of a single record on the analysis results to protect individual privacy.

Definition 1 (*Differential Privacy [16]*). Let $\varepsilon \geqslant 0$ and $\delta \in [0,1)$. A randomized mechanism $\mathscr{M}: \mathscr{X}^n \to \mathscr{Y}$satisfies (ε, δ)-DP if for any two datasets $X, X' \in \mathscr{X}^n$ that differ in only one record, and any $Y \subseteq \mathscr{Y}$, it holds that

$$Pr[\mathscr{M}(X) \in Y] \leqslant e^{\varepsilon} Pr[\mathscr{M}(X') \in Y] + \delta.$$

The effect of a single record on the analysis results is limited by ε and δ. The parameter ε restricts the difference between two output distributions. A larger ε indicates a higher privacy risk. δ is usually a negligible value to count the part of the output distributions differences that exceed ε.

2.2. Mechanisms in Differential Privacy

We introduce two popular DP mechanisms: the Laplace and the Gaussian mechanisms. When computing function f, two randomized mechanisms add a random noise drawn from the Laplace or Gaussian distributions to satisfy DP. The scale of the noise is regulated by the global sensitivity Δf.

[1] https://github.com/con-fide/TVDprivacy

Definition 2 (*Global Sensitivity [16]*). Given any function $f : \mathscr{X}^n \to \mathbb{R}^d$, for any two datasets $X, X' \in \mathscr{X}^n$ that differ in only one record, the global sensitivity of function f is

$$\Delta_p f = \max_{X,X'} ||f(X) - f(X')||_p,$$

where $||\cdot||_p$ denotes the l_p norm.

Theorem 1 (*Laplace Mechanism [16]*). Given any function $f : \mathscr{X}^n \to \mathbb{R}^d$, the Laplace mechanism is defined as $\mathscr{M}_L(X) = f(X) + (\eta_1, \cdots, \eta_d)$, which satisfies $(\varepsilon, 0)$-DP, where η_i denotes random variables independently drawn from $\mathrm{Lap}\left(\frac{\Delta_1 f}{\varepsilon}\right)$.

Theorem 2 (*Gaussian Mechanism [16]*). Given any function $f : \mathscr{X}^n \to \mathbb{R}^d$, the Gaussian mechanism is defined as $\mathscr{M}_G(X) = f(X) + (\eta_1, \cdots, \eta_d)$, which satisfies (ε, δ)-DP, where η_i denotes random variables independently drawn from $N\left(0, \Delta_2^2 f \sigma^2\right)$ with $\sigma = \frac{\sqrt{2\log(1.25/\delta)}}{\varepsilon}$.

2.3. Total Variation Distance

Total Variation Distance (TVD, also called the statistical difference or the statistical distance) is a distance measurement between two probability distributions. For two probability measures P and Q on a random variable X, the TVD measure between P and Q is defined as

$$D_{TV}(P, Q) = \max_{X \subset \mathscr{X}} |P(X) - Q(X)|.$$

2.4. Kullback–Leibler Divergence

Kullback–Leibler (KL) divergence (also called relative entropy) is a measure of the difference between two probability distributions. For discrete probability distribution P, Q in the probability space $\mathbb{X}$, KL divergence $D_{KL}(P||Q)$ is defined to be

$$D_{KL}(P||Q) = \sum_{x \in \mathbb{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

For continuous probability distribution $P, Q, D_{KL}(P||Q)$ is defined as

$$D_{KL}(P||Q) = \int_{-\infty}^{+\infty} P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx.$$

With the KL divergence, we can calculate the upper bound of TVD. For two probability distribution P, Q, if we have the KL divergence $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$, we can calculate the upper bound of $D_{TV}(P, Q)$ as follows:

$$D_{TV}(P, Q) \leqslant max\left[\sqrt{\tfrac{1}{2} D_{KL}(P||Q)}, \sqrt{\tfrac{1}{2} D_{KL}(Q||P)}\right].$$

When $D_{KL}(P||Q) \geqslant 2$ or $D_{KL}(Q||P) \geqslant 2$, we can calculate the upper bound of TVD privacy as follows:

$$D_{TV}(P, Q) \leqslant max\left[\sqrt{1 - e^{-D_{KL}(P||Q)}}, \sqrt{1 - e^{-D_{KL}(Q||P)}}\right].$$

2.5. Inference Attacks

In an inference attack, the adversary usually uses the output of the data analysis task and the background knowledge to infer the attribute, category, or other sensitive information about a given record. In recent years, some inference attacks [1–5] started to use training models to expose sensitive information of training samples. Some inference attacks do not require strong assumptions and can infer some sensitive information with a high probability just through a black-box API.

In this paper, we use membership inference attacks (MIAs) as an example to analyze the ability of privacy definition to estimate privacy risks. In machine learning, MIAs aim to infer whether a given record is a training sample. Shokri et al. [1] first proposed MIAs and assumed that the adversary has black-box model access and a given dataset. They successfully implemented their attacks on services and suggested using MIA as a metric for choosing a training model. We use the membership experiment proposed by Yeom et al. [9] to formalize the MIAs.

Experiment 1. (*Membership experiment* $Exp^M(\mathscr{A}, \mathscr{M}, n, D)$[9]). Let $\mathscr{A}$ be an adversary, $\mathscr{M}$ be a learning algorithm, n be a positive integer, and D be a distribution over data points (x, y). The membership experiment proceeds as follows:

(1)Sample $S \sim D_n$ and let $\mathscr{M}_S = \mathscr{M}(S)$.
(2)Choose $b \leftarrow \{0,1\}$ uniformly at random.
(3)Draw $z \sim S$ if $b = 0$ or $z \sim D$ if $b = 1$.
(4)$Exp^M(\mathscr{A},\mathscr{M},n,D)$ is 1 if $\mathscr{A}(z,\mathscr{M}_S,n,D) = b$ and 0 otherwise. $\mathscr{A}$ must output either 0 or 1.

3. Differential Privacy Performance on Estimating Inference Attacks

DP parameters (ε,δ) measure the randomness of private mechanisms in the worst case. However, when users aim to estimate the privacy risks for some specific attacks, these parameters sometimes do not provide an accurate estimation. We explain this issue with an inference experiment.

Experiment 2. (*Inference experiment* $Exp^I(\mathscr{A},\mathscr{M},X_0,X_1)$) Let $\mathscr{A}$ be an adversary, $\mathscr{M}$ be a randomized mechanism, and X_0 and X_1 be two possible inputs. The inference experiment proceeds as follows:

(1) Choose $b \leftarrow \{0,1\}$ such that $b = 0$ with probability P_0 and $b = 1$ with probability P_1.
(2) Compute $y = \mathscr{M}(X_b)$.
(3) $Exp^I(\mathscr{A},\mathscr{M},X_0,X_1)$ is 1 if $\mathscr{A}(y,\mathscr{M},X_0,X_1) = b$ and 0 otherwise. $\mathscr{A}$ must output either 0 or 1.

Let X_0 and X_1 correspond to inputs 0 and 1 and $P_0 = P_1 = 1/2$. Suppose we can use the Laplace or Gaussian mechanisms to protect privacy. These two mechanisms add the noise drawn from $Lap(1)$ and $N(0,2)$ to the output. We first show how an adversary would design its inference strategy. The adversary has the balance prior probabilities on the input. It needs to design an inference strategy to maximize the inference accuracy, i.e., to maximize $E\left(Exp^I\right)$. We assume that the adversary infers $\mathscr{A}(y) = 0$ for any $y \in y_0$ and infers $\mathscr{A}(y) = 1$ for any $y \in y_1$, where y_0 and y_1 are complementary. $E\left(Exp^I\right)$ is calculated as follows:

$$\begin{aligned} E(Exp^I) &= \int_{y_0} P_0 \cdot Pr[\mathscr{A}(y) = 0|X = 0]dy + \int_{y_1} P_1 \cdot Pr[\mathscr{A}(y) = 1|X = 1]dy \\ &= P_0 \cdot Pr[\mathscr{M}(0) \in y_0] + P_1 \cdot Pr[\mathscr{M}(1) \in y_1] \\ &= \frac{1}{2}(Pr[\mathscr{M}(0) \in y_0] + Pr[\mathscr{M}(1) \in y_1]). \end{aligned}$$

To maximize $E\left(Exp^I\right)$, the adversary infers $\mathscr{A}(y) = 0$ when $Pr[\mathscr{M}(0) = y] > Pr[\mathscr{M}(1) = y]$ and infers $\mathscr{A}(y) = 1$ when $Pr[\mathscr{M}(0) = y] \leqslant Pr[\mathscr{M}(1) = y]$. Using this inference strategy, the adversary has a correct probability of 0.697 for the Laplace mechanism and a correct probability of 0.639 for the Gaussian mechanism.

However, DP provides the opposite suggestion. The Laplace mechanism is $(1,0)$-DP, and the Gaussian mechanism is $\left(3.4,10^{-5}\right)$-DP. Under the metric of DP, the Laplace mechanism is more private than the Gaussian mechanism. For any (ε,δ)-DP mechanism, the upper bound of $E\left(Exp^I\right)$ [2] is

$$\begin{aligned} E(Exp^I) &= P_0 \cdot Pr[\mathscr{M}(0) \in y_0] + P_1 \cdot Pr[\mathscr{M}(1) \in y_1] \\ &\leq \frac{1}{2}\left(\frac{e^\varepsilon}{e^\varepsilon+1}(1-\delta)+\delta\right) + \frac{1}{2}\left(\frac{e^\varepsilon}{e^\varepsilon+1}(1-\delta)+\delta\right) \\ &= \frac{e^\varepsilon}{e^\varepsilon+1}(1-\delta)+\delta. \end{aligned} \tag{1}$$

Therefore, the upper bound of $E\left(Exp^I\right)$ of the Gaussian mechanism is 0.97, and that of the Laplace mechanism is 0.73.

In Experiment 2, we should choose the Gaussian mechanism to limit the accuracy of the inference attack; however, DP does not provide the correct suggestion. Assume we can only use the Gaussian mechanism and limit the inference accuracy to no more than 0.639. We only need to add noise with $\sigma = 2$ instead of the DP-guided noise with $\sigma > 7$. To reduce noise, we need an inference measurement with a more accurate estimate of the privacy risk.

4. Total Variation Distance Privacy

In this section, we formally introduce TVD privacy and analyze the advantages and disadvantages of TVD privacy in estimating privacy risks.

[2] The bound in Eq. 1 has appeared in [13,15].

Definition 3. A randomized mechanism $\mathscr{M}: \mathscr{X}^n \to \mathscr{Y}$satisfies α-TVD privacy if for any two datasets $X, X' \in \mathscr{X}^n$ that differ in only one record, it holds

$$\sup_{Y \subseteq \mathscr{Y}} |Pr[\mathscr{M}(X) \in Y] - Pr[\mathscr{M}(X') \in Y]| \leqslant \alpha.$$

TVD privacy strictly limits the difference between the output distributions of private mechanisms. In Experiment 2, the adversary inference function is used to determine the output set, which maximizes the TVD privacy parameter α. Then the maximum accuracy of inference attacks is equal to $1/2 + 1/2\alpha$. In Property 1, we further analyze the guarantee of TVD privacy, which can also provide an accurate measure for unbalanced prior probabilities.

4.1. TVD Privacy Guarantee

TVD privacy ensures that the output distribution is roughly the same whether the record is in the dataset. In this aspect, TVD privacy and DP are similar. With the guarantee of TVD privacy, users know that the probabilities of privacy leakages are approximately the same whether they participate in data collection.[3]

TVD privacy can accurately estimate the accuracy of inference attacks. In particular, TVD privacy provides the most accurate estimation when the prior probabilities of neighboring datasets are the same. We analyze the measure of TVD privacy with arbitrary prior probabilities for Experiment 2.

Proposition 1. For any prior probability $P_0, P_1 (P_0 \geqslant P_1)$ on two datasets X_0, X_1 that differ in only one record, if we use an α-TVD private mechanism and the adversary has the prior probabilities P_0, P_1, the accuracy of adversary inference is not more than $P_0 + P_1\alpha$. When $P_0 = P_1$, the accuracy is not more than $1/2 + 1/2\alpha$.

Proof 1.

Let $Y \in \mathscr{Y}$ maximize $Pr[\mathscr{M}(X_0) \in Y] - Pr[\mathscr{M}(X_1) \in Y]$. We set $Pr[\mathscr{M}(X_0) \in Y] = a + \alpha$ and $Pr[\mathscr{M}(X_1) \in Y] = a$ for any $a \in [0, (1-\alpha)/2]$. The adversary designs the inference strategy with the prior probability P_0, P_1. The adversary can design a set S to maximize $E\left(Exp^I\right)$. Let $\mathscr{A}(y) = 0$ when $y \in S$ and $\mathscr{A}(y) = 1$ when $y \notin S$. The maximum $E\left(Exp^I\right)$ can be expressed as follows:

$$\begin{aligned} E(Exp^I) &= P_0 \cdot Pr[\mathscr{M}(X_0) \in S] + P_1 \cdot Pr[\mathscr{M}(X_1) \notin S] \\ &= P_0 \cdot Pr[\mathscr{M}(X_0) \in Y] + P_0 \cdot Pr[\mathscr{M}(X_0) \in (S-Y)] + P_1 \cdot Pr[\mathscr{M}(X_1) \notin Y] - P_1 \cdot Pr[\mathscr{M}(X_1) \in (S-Y)]. \end{aligned} \tag{2}$$

Let $Pr[\mathscr{M}(X_0) \in (S-Y)] = b_0$ and $Pr[\mathscr{M}(X_1) \in (S-Y)] = b_1$; then we have (2)

$$= P_0(a+\alpha) + P_0 b_0 + P_1(1-a) - P_1 b_1 \tag{3}$$

Since $b_0 \leqslant 1 - a - \alpha$ and $b_1 \geqslant b_0$, we have(3)

$$\leqslant P_0\alpha + (P_0 - P_1)a + P_1 + (P_0 - P_1)(1 - a - \alpha) = P_0 + P_1\alpha. \tag{4}$$

The proof is complete.

4.2. Difference between TVD Privacy and Differential Privacy

We take the Gaussian mechanism as an example to explain the difference between TVD privacy and DP. In Experiment 2, inputs 0 and 1 correspond to output distributions $N(0,1)$ and $N(1,1)$, respectively. Fig. 1 illustrates the probability density functions of $N(0,1)$ and $N(1,1)$. DP requires that the ratio of two probability density functions be close. Therefore, the privacy loss is insignificant, regardless of the mechanism output. TVD privacy is not as strict as DP. It quantifies the area of the shaded portion as a privacy loss. That is, $Pr[\mathscr{M}(0) < 0.5] - Pr[\mathscr{M}(1) < 0.5]$ should be small enough.

We use hypothesis testing to help the reader understand the relationship between actual privacy guarantees, DP, and TVD privacy. In Experiment 2, the adversary obtains the output y of the private mechanism. It chooses a null hypothesis of H_0 and an alternative hypothesis of H_1:

H_0: y came from input 0,

H_1: y came from input 1.

[3] TVD privacy adheres to the same privacy philosophy as DP, i.e., an individual's control over information is the target of privacy protection. The promise of differential privacy versus TVD privacy does not cover information leakage due to data correlations. A series of works have discussed this point [13,17–25]. At the heart of this issue is what we should protect as private information. Views on information privacy can be divided into control over personal information and limited access to personal information. We briefly explore the two dominant privacy philosophies in A and hope that this part will help the reader better understand privacy instead and limit misconceptions.

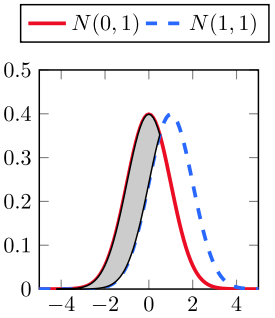

Fig. 1. TVD privacy and differential privacy for Gaussian mechanism.

The adversary chooses the rejection region S, which corresponds to the false negative rate as $P_{FN}(\mathscr{M}, S) = Pr[\mathscr{M}(0) \in S]$. Let $\overline{S}$ denote the complement of S. Then, the false positive probability is $P_{FP}(\mathscr{M}, S) = Pr\left[\mathscr{M}(1) \in \overline{S}\right]$. Kairouz proved the following relationship between DP mechanisms and hypothesis testing.

Theorem 3. As stated in [26], for any $\alpha > 0$ and $\delta \in [0, 1]$, a randomized mechanism $\mathscr{M}$ is (α, δ)-DP if and only if for any two neighboring databases X_0, X_1 and any rejection region S, it holds

$$Pr[\mathscr{M}(X_0) \in S] + e^{\varepsilon} Pr\left[\mathscr{M}(X_1) \in \overline{S}\right] \geqslant 1 - \delta,$$
$$e^{\varepsilon} Pr[\mathscr{M}(X_0) \in S] + Pr\left[\mathscr{M}(X_1) \in \overline{S}\right] \geqslant 1 - \delta.$$

Correspondingly, we formally provide the relationship between TVD privacy and hypothesis testing. TVD privacy limits the lower bound of the sum of the false positive rate and false negative rate.

Theorem 4. For any $\alpha > 0$, a randomized mechanism $\mathscr{M}$ satisfies α-TVD privacy if and only if for any two datasets X, X' that differ in only one record, and any rejection region S, it holds

$$Pr[\mathscr{M}(X_0) \in S] + Pr\left[\mathscr{M}(X_1) \in \overline{S}\right] \geqslant 1 - \alpha.$$

TVD privacy guarantees that for any rejection region S, the sum of the false positive rate and the false negative rate is not less than $1 - \alpha$.

In Fig. 2, we plot the actual curve of the Gaussian mechanism, the DP curve, and the TVD privacy curve. Both DP and TVD privacy are descriptions of the actual privacy curve. We can view DP and TVD privacy as two complementary privacy definitions. DP is more accurate than TVD privacy when the false negative rate is approaching the limit. In contrast, TVD privacy is more accurate when the sum of the false positive and false negative rates is large. From Theorem 3 and Theorem 4, we can calculate that TVD privacy is more accurate in $\left(\frac{\alpha-\delta}{e^{\varepsilon}-1}, \frac{e^{\varepsilon}(1-\alpha)-1-\delta}{e^{\varepsilon}-1}\right)$, and DP is more accurate outside of this interval.

4.3. Formal Properties for TVD Privacy

We provide several formal properties for TVD privacy. These properties guarantee the availability of TVD privacy in realistic scenarios. First, we prove that TVD privacy satisfies the post-processing property. The post-processing property ensures that a subsequent analyses of the randomized mechanism output does not weaken the TVD privacy guarantee if no additional dataset information is available.

Theorem 5. (***Post-processing Theorem***). Let $\mathscr{M} : \mathscr{X}^n \to \mathscr{Y}$ be a randomized mechanism satisfing α-TVD privacy and $f : \mathscr{Y} \to \mathscr{Z}$ be any arbitrary randomized mapping. Then $f \circ \mathscr{M} : \mathscr{X}^n \to \mathscr{Z}$ satisfies α-TVD privacy.

Proof 2.

For any two datasets $X, X' \in \mathscr{X}^n$ that differ in only one record, any event $Z \subset \mathscr{Z}$. Let $Y = \{y \in Y : f(y) \in Z\}$. We have

$$Pr[f(\mathscr{M}(X)) \in Z] = Pr[\mathscr{M}(X) \in Y] = Pr[\mathscr{M}(X') \in Y] + \alpha = Pr[f(\mathscr{M}(X')) \in Z] + \alpha.$$

The proof is complete.

Next, we provide the composition theorem of TVD privacy. The composition theorem guarantees the availability of TVD privacy for multiple queries on a dataset. We first provide a naive composition theorem for TVD privacy, which provides a lower bound for any TVD algorithm.

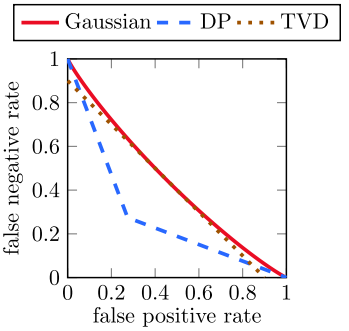

Fig. 2. Hypothesis testing curve.

Theorem 6 (***Naive Composition Theorem***). Let $\mathscr{M}_1 : \mathscr{X} \to \mathscr{Y}_1$ be an α_1-TVD private mechanism, and let $\mathscr{M}_2 : \mathscr{X} \to \mathscr{Y}_2$ be an α_2-TVD private mechanism. Then their composition, defined as $\mathscr{M}_{1,2} : \mathscr{X} \to \mathscr{Y}_1 \times \mathscr{Y}_2$ by the mapping: $\mathscr{M}_{1,2}(X) = (\mathscr{M}_1(X), \mathscr{M}_2(X))$ is $(\alpha_1 + \alpha_2 - \alpha_1 \times \alpha_2)$-TVD privacy.

Proof 3.

Since $\mathscr{M}_1, \mathscr{M}_2$ satisfy α_1- and α_2-TVD privacy, respectively. At least $1 - \alpha_1$ of the distributions of $\mathscr{M}_1(X)$ and $\mathscr{M}_1(X')$ overlap. Similarly, at least $1 - \alpha_2$ of the distributions of $\mathscr{M}_2(X)$ and $\mathscr{M}_2(X')$ are overlap. Therefore, at least $(1 - \alpha_1)(1 - \alpha_2)$ of the distributions of $\mathscr{M}_1(X) \cdot \mathscr{M}_2(X)$ and $\mathscr{M}_1(X') \cdot \mathscr{M}_2(X')$ are overlapped. The TVD distance between $\mathscr{M}_{1,2}(X)$ and $\mathscr{M}_{1,2}(X')$ is no more than $1 - (1 - \alpha_1)(1 - \alpha_2) = \alpha_1 + \alpha_2 - \alpha_1 \times \alpha_2$. The proof is complete.

If we have the KL divergence of randomized mechanisms, denoted as $D_{KL}(\mathscr{M}(X)||\mathscr{M}(X'))$, we can provide a more precise composition theorem. First, we define the KL divergence of the randomized mechanism $D_{KL}(\mathscr{M})$.

Definition 4. For any two datasets X, X' that differ in only one record, KL divergence of mechanism $\mathscr{M}$ is:

$$D_{KL}(\mathscr{M}) \geqslant \sup_{X,X'} max\{D_{KL}(\mathscr{M}(X)||\mathscr{M}_1(X')), D_{KL}(\mathscr{M}(X')||\mathscr{M}(X))\}.$$

Theorem 7. (***Advanced Composition Theorem***). Let $\mathscr{M}_1 : \mathscr{X}^n \to \mathscr{Y}_1$ and $\mathscr{M}_2 : \mathscr{X}^n \to \mathscr{Y}_2$ be two randomized mechanisms. If we have the KL divergence of $\mathscr{M}_1$ and $\mathscr{M}_2$, denote as $D_{KL}(\mathscr{M}_1), D_{KL}(\mathscr{M}_2)$, then their composition, $\mathscr{M}_{1,2}(X) = (\mathscr{M}_1(X), \mathscr{M}_2(X))$, satisfies $\sqrt{1/2(D_{KL}(\mathscr{M}_1) + D_{KL}(\mathscr{M}_2))}$-TVD privacy.

Proof 4.

Because KL divergence is additive for independent distributions, then we have

$$D_{KL}(\mathscr{M}_{1,2}) \leqslant D_{KL}(\mathscr{M}_1) + D_{KL}(\mathscr{M}_2).$$

Using Pinsker's inequality, we have

$$\alpha \leqslant \sqrt{1/2 D_{KL}(\mathscr{M}_{1,2})} \leqslant \sqrt{1/2(D_{KL}(\mathscr{M}_1) + D_{KL}(\mathscr{M}_2))}.$$

The proof is complete.

When $D_{KL}(\mathscr{M}) > 2$, we can use the following inequality proposed by Bretagnolle and Huber to compute α to ensure that α is always less than 1:

$$\alpha \leqslant \sqrt{1 - e^{-D_{KL}(\mathscr{M})}}.$$

Next, we provide the parallel composition property of TVD privacy. The parallel composition property implies that each record has privacy risk only when it is used. When a portion of the dataset is queried, it does not compromise the privacy of other records.

Theorem 8. (***Parallel Composition Theorem***). Let $\mathscr{M}_i$ sastisfy α_i-TVD privacy, X_i be disjoint subsets of X, and $X_1 \cup \cdots \cup X_n = X$. The algorithm $\mathscr{M} = (\mathscr{M}_1(X_1), \cdots, (\mathscr{M}_n(X_n))$ satisfies $max(\alpha_i)$-TVD privacy.

Proof 5.

For two datasets X, X' that differ in only one record, X, X' are divided into disjoint n subsets $(X_1, \cdots, X_n)$ and $(X_1\prime, \cdots, X_n\prime)$. Without loss of generality, we assume that $X_i = X_i\prime$ for any $i \neq j$, and X_j, X_j' are different in only one record. Then, we have

$$\begin{aligned} Pr[\mathscr{M}(X) \in \{Y_1, \cdots, Y_n\}] &= \prod Pr[\mathscr{M}(X_i) \in Y_i] \\ &\leqslant \left(Pr\left[\mathscr{M}\left(X_j'\right) \in Y_j\right] + \alpha_j\right) \prod_{i=2}^{n} Pr[\mathscr{M}(X_i) \in y_i] \\ &\leqslant Pr\left[\mathscr{M}\left(X_j'\right) \in Y_j\right] \prod_{i \neq j} Pr[\mathscr{M}(X_i\prime) \in Y_i] + \alpha_j \\ &= Pr[\mathscr{M}(X') \in \{Y_1, \cdots, Y_n\}] + \alpha_j. \end{aligned}$$

Therefore, for any two datasets X and X' that differ in only one record, the privacy loss of $\mathscr{M}$ does not exceed $max(\alpha_i)$.

Then, we turn our attention to post-processing for privacy enhancement. In DP, some studies focus on providing privacy guarantees through post-processing noise [27,28]. In particular, Erlingsson et al. provided a general result for any ε-DP mechanism [29]. We provide a general theorem for TVD privacy.

Theorem 9 (*Chained Composition Theorem*). For any α_1-TVD private algorithm $\mathscr{M}_1 : \mathscr{X} \to \mathscr{Y}$ and any α_2- TVD private algorithm $\mathscr{M}_1 : \mathscr{Y} \to \mathscr{Z}$, we have that $\mathscr{M}_2 \circ \mathscr{M}_1$ is a $(\alpha_1\alpha_2)$-TVD private algorithm.

Proof 6.

Since $\mathscr{M}_1$ satisfies α_1-TVD privacy, for any two datasets X, X' and $Z \subset \mathscr{Z}$, there must be a set $Y \in \mathscr{Y}$ that maximizes $Pr[\mathscr{M}_1(X) \in Y] - Pr[\mathscr{M}_1(X') \in Y]$. Let $\overline{Y}$ be the complement of Y and $\overline{Z}$ be the complement of Z. We have

$$\begin{aligned} Pr[\mathscr{M}_2(\mathscr{M}_1(X)) \in Z] - Pr[\mathscr{M}_2(\mathscr{M}_1(X')) \in Z] &= \sum_{y \in Y} Pr[\mathscr{M}_1(X) = y] Pr[\mathscr{M}_2(y) \in Z] \\ &+ \sum_{y \in \overline{Y}_1} Pr[\mathscr{M}_1(X) = y] Pr[\mathscr{M}_2(y) \in Z] \\ &- \sum_{y \in Y} Pr[\mathscr{M}_1(X') = y] Pr[\mathscr{M}_2(y) \in Z] \\ &- \sum_{y \in \overline{Y}_1} Pr[\mathscr{M}_1(X') = y] Pr[\mathscr{M}_2(y) \in Z] \\ &= \sum_{y \in Y} (Pr[\mathscr{M}_1(X) = y] - Pr[\mathscr{M}_1(X') = y]) Pr[\mathscr{M}_2(y_1) \in Z] \\ &- \sum_{y \in \overline{Y}} (Pr[\mathscr{M}_1(X') = y] - Pr[\mathscr{M}_1(X) = y]) Pr[\mathscr{M}_2(y_1) \in Z]. \end{aligned} \tag{5}$$

Since $\mathscr{M}_2$ also satisfies α_2-TVD privacy, we have $Pr[\mathscr{M}(y) \in Z] - Pr[\mathscr{M}(y\prime) \in Z] \leqslant \alpha_2$. Let $p = max\{Pr[\mathscr{M}(y) \in Z]\}$; then, we have (5)

$$\leqslant \alpha_1 p - \alpha_1(p - \alpha_2) = \alpha_1\alpha_2.$$

The proof is complete.

Similar to the group privacy of DP [16], we also provide group privacy of TVD privacy to measure the privacy leakage of a group of records in the data analysis tasks.

Definition 5 (*Group privacy of TVD privacy*). A randomized mechanism $\mathscr{M} : \mathscr{X}^n \to \mathscr{Y}$satisfies α-TVD privacy if for any $X, X' \in \mathscr{X}^n$ that differ in k element, it holds

$$\sup_{Y \subseteq \mathscr{Y}} |Pr[\mathscr{M}(X) \in Y] - Pr[\mathscr{M}(X') \in Y]| \leqslant \alpha.$$

Similar to the group privacy of DP, the strength of the privacy guarantee of group privacy of TVD privacy decreases with the size of the group.

4.4. Privacy Amplification by Sampling

The privacy definition can benefit from the uncertainty of the adversary over the dataset [30–33]. We analyze the TVD privacy amplification by sampling. Assume the data are sampled from dataset X. The adversary aims to distinguish two datasets X and X', that differ in only one record. Suppose the record x_n is absent in the dataset X' and the sampling rate is q. The privacy amplification is linearly related to the sampling rate q.

Theorem 10. For any mechanism $\mathscr{M}$ that satisfies α-TVD privacy, if the input is sampled from the dataset X with a probability q, denoted as $S_q(X)$, then $\mathscr{M} \circ S_q$ satisfies $(q\alpha)$-TVD privacy.

See B for the proof.

4.5. TVD Private Mechanisms

In this section, we show how to use TVD privacy to analyze private mechanisms. We analyze the Laplace and Gaussian mechanisms in single-dimensional numerical queries and the Gaussian mechanism in multi-dimensional numerical queries. These private mechanisms can solve a large number of statistical query tasks.

4.5.1. Single-Dimensional Mechanisms

We use TVD privacy to measure the Laplace and Gaussian mechanisms in a single-dimensional numerical query. The sensitivity of the counting query is $\Delta_1 f = \Delta_2 f = 1$.

Theorem 11 (*Laplace Mechanism for TVD Privacy*). Given any function $f : \mathscr{X}^n \to \mathbb{R}$, the Laplace mechanism defined as

$$\mathscr{M}_L(X) = f(X) + \eta,$$

which satisfies $\left(1 - e^{-\frac{\Delta f}{2b}}\right)$-TVD privacy, where η is a random variable drawn from $Lap(b)$.

Proof 7.

Let μ denote the mean of the Laplace distribution. The Laplace distribution has a cumulative distribution function (CDF) $F_L(x|\mu, b)$, and we can compute TVD privacy as follows:

$$D_{TV}(\mathscr{M}(X)||\mathscr{M}(X')) \leq F_L(\frac{1}{2}\Delta f|0, b) - F_L(\frac{1}{2}\Delta f|\Delta f, b) = \frac{1}{2}(1 - e^{-\frac{|\Delta f|}{2b}}) + \frac{1}{2}(1 - e^{-\frac{|-\Delta f|}{2b}}) = 1 - e^{-\frac{\Delta f}{2b}}.$$

The proof is complete.

Theorem 12 (*Gaussian Mechanism for TVD Privacy*). Given any function $f : \mathscr{X}^n \to \mathbb{R}$, the Gaussian mechanism is defined as

$$\mathscr{M}_G(X) = f(X) + \eta,$$

which satisfies $\left(2\Phi\left(\frac{\Delta f}{2\sigma}\right) - 1\right)$-TVD privacy, where η is a random variable drawn from $N(0, \sigma^2)$ and $\Phi(x)$ denotes the CDF of the standard normal distribution.

Proof 8.

Let $\Phi(x)$ denote the CDF of the standard normal distribution, and μ denote the mean of the normal distribution. Assuming that the Gaussian mechanism adds noise drawn from $N(0, \sigma^2)$, the CDF of the Gaussian distribution is $F_G(x|\mu, \sigma^2)$. We have

$$D_{TV}(\mathscr{M}(X)||\mathscr{M}(X')) \leq F_G(\frac{1}{2}\Delta f|0, \sigma^2) - F_G(\frac{1}{2}\Delta f|\Delta f, \sigma^2) = 2F_G(\frac{1}{2}\Delta f|0, \sigma^2) - 1 = 2\Phi\left(\frac{\Delta f}{2\sigma}\right) - 1.$$

The proof is complete.

4.5.2. Multi-Dimensional Gaussian Mechanism

We prove the TVD privacy of the Gaussian mechanism in multi-dimensional numerical queries. The proof is similar to the proof of the multi-dimensional Gaussian mechanism in [16].

Theorem 13 (*Multi-dimensional Gaussian Mechanism for TVD Privacy*). Given any function $f : \mathscr{X}^n \to \mathbb{R}^d$, the Gaussian mechanism is defined as

$$\mathscr{M}_G(X) = f(X) + (\eta_1, \cdots, \eta_d),$$

which satisfies $\left(2\Phi\left(\frac{\Delta_2 f}{2\sigma}\right) - 1\right)$-TVD privacy, where η_i are random variables independently drawn from $N(0, \sigma^2)$.

Proof 9.

Since the Gaussian distribution is independent of the orthogonal basis, we can transform a multi-dimensional Gaussian mechanism into a single-dimensional Gaussian mechanism. Consider query function $f(X) = (y_1, \cdots, y_m)$ with sensitivity $\Delta_2 f$. For any two datasets X, X' that differ in only one record, let $\mathbf{v} = f(X) - f(X')$; then, we have $\|\mathbf{v}\|_2 \leqslant \Delta_2 f$.

Since the multi-dimensional Gaussian distribution is spherically symmetric, we can choose another standard orthogonal basis $(b_1, \cdots, b_m)$ where b_1 is parallel to $\mathbf{v}$. The Gaussian mechanism adds noise drawn from $N(0, \sigma^2)$ in the direction of $\mathbf{v}$. For another basis $b_i, f(X)$ and $f(X')$ are identical. For any two datasets X, X' that differ in only one record, we have

$$D_{TV}(\mathscr{M}(X)||\mathscr{M}(X')) = D_{TV}(\mathscr{M}(X)||\mathscr{M}(X)+\mathbf{v}) \le F_G\left(\frac{1}{2}\Delta_2 f|0,\sigma^2\right) - F_G\left(\frac{1}{2}\Delta_2 f|\Delta_2 f,\sigma^2\right) = 2F_G\left(\frac{1}{2}\Delta_2 f|0,\sigma^2\right) - 1 = 2\Phi\left(\frac{\Delta_2 f}{2\sigma}\right) - 1.$$

The proof is complete.

5. TVD Privacy for Membership Inference Attacks

Membership inference attacks are familiar attacks on training samples in machine learning. It aims to infer whether a record is a training sample or not. Jayaraman et al. [10] mentioned that MIAs are the most directly related to DP among attacks. Similarly, MIAs are directly related to the protection target of TVD privacy. The study of privacy definitions and MIAs helps to understand the potential privacy risk in data analysis tasks. In this section, we show the advantage of TVD privacy in estimating the privacy risk of MIAs.

5.1. Private Stochastic Gradient Descent

In machine learning, Stochastic Gradient Descent (SGD) is commonly used to minimize the loss function L. A general approach to protect the privacy of training samples is to use a private SGD algorithm. We analyze the TVD privacy property of the differentially private SGD (DP-SGD) algorithm [6]. The DP-SGD algorithm randomly selects a small batch of training samples in each training step. Before updating each gradient, DP-SGD adds noise drawn from a Gaussian distribution to protect the privacy of training samples.

First, DP-SGD restricts the boundary of each sample gradient via the clipping bound C. Then, DP-SGD adds Gaussian noise drawn from $N\left(0, \Delta_2^2 f \sigma^2\ \mathbb{I}\right)$. For each step, the update rule is

$$\theta_{i+1} \leftarrow \theta_i - \eta \frac{1}{b}\left(\sum g_t(x_i) + N\left(0, \Delta_2^2 f \sigma^2 \mathbb{I}\right)\right),$$

where η denotes the learning rate, b denotes the number of samples drawn from each batch, θ_i denotes the gradient in this round, and $g_t(x_i)$ denotes the update gradient corresponding to sample x_i. Let the number of training samples be n. Each training sample is sampled in each step with probability $q = b/n$. Thus, the privacy guarantee of DP-SGD is equivalent to that of the sampled Gaussian mechanism $\mathscr{M}_{SG}$ [7]. Fortunately, $\mathscr{M}_{SG}$ can be similarly reduced to a single-dimensional sampled Gaussian mechanism [7]. Therefore, we can conveniently use the advanced composition theorem to analyze the TVD privacy for $\mathscr{M}_{SG}$.

The Killback-Leibler divergence of the sampled Gaussian distribution is

$$D_{KL}(\mathscr{M}_{SG}(\sigma,q)) = max\{D_{KL}\left(N(0,\sigma^2)||(1-q)N(0,\sigma^2) + qN(1,\sigma^2)\right), D_{KL}\left((1-q)N(0,\sigma^2) + qN(1,\sigma^2)||N(0,\sigma^2)\right)\}.$$

By Theorem 7, DP-SGD satisfies $\sqrt{\frac{1}{2}\ T \cdot D_{KL}(\mathscr{M}_{SG}(\sigma,q))}$-TVD privacy.

5.2. The Accuracy of MIAs

In MIAs [1,9,11], a typical setup is that half of the target samples are training samples and the others are not. In this setup, an adversary has the same prior probability for the target sample. The accuracy is equal to the ratio of correct predictions among all predictions.

First, we analyze which of TVD privacy or DP can more accurately measure the upper bound of the accuracy of MIA. Assume that DP-SGD algorithm satisfies (ε, δ)-DP and α-TVD privacy. From Eq. 4, TVD privacy guarantees that the accuracy of an MIA is no more than $(1/2 + 1/2\alpha)$. From Eq. 1, DP promises that the accuracy of MIA is less than $\left(\frac{e^\varepsilon}{1+e^\varepsilon}(1-\delta)+\delta\right)$. The privacy property of DP-SGD is related to the sampling probability q, the number of steps T, and the Gaussian noise parameter σ. Fig. 3 shows the upper bounds of the accuracy of MIA for different parameters. As the theoretical analysis in Section 4.1, TVD privacy always gives a more accurate theoretical upper bound than DP for the accuracy of MIAs.

Second, we compare the training results of TVD privacy and DP. When we try to constrain the accuracy of MIAs not to exceed a certain threshold, TVD privacy and DP can set the scale of noise based on their respective theoretical analyses. We consider four cases from low to high accuracy of MIAs and require the private SGD algorithms to guarantee that the accuracy of MIAs does not exceed $0.8, 0.7, 0.6$, and 0.55, respectively. With the privacy analysis method of DP-SGD in Section 5.1, Eqs. 1 and 4, we can compute the noise scales for TVD privacy and DP. We train with the tutorial code provided by tensorflow [4]. Fig. 4 shows the training accuracy for TVD privacy and DP. At a low privacy level (with an MIA's accuracy of 0.8 or 0.7), although TVD privacy adds less noise than DP, both have high training accuracy. However, at a high privacy level (with an MIA's accuracy of 0.6 or 0.55), the training accuracy of TVD privacy is significantly higher than that of DP.

[4] https://github.com/tensorflow/privacy/tree/master/tutorials/mnist_dpsgd_tutorial.py

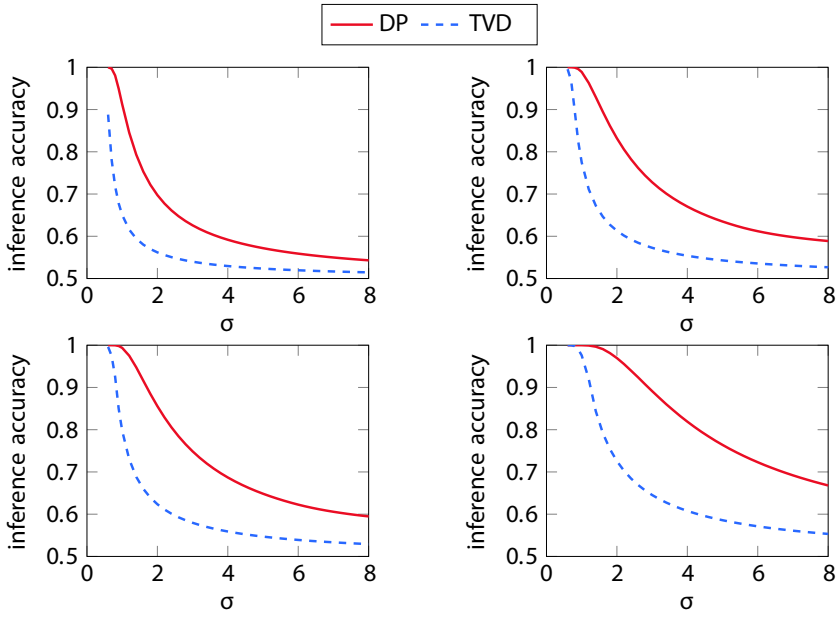

Fig. 3. The upper bound on the accuracy of MIAs. The batch size and steps are set at $(300, 6000)$, $(300, 20000)$, $(600, 6000)$ and $(600, 20000)$, respectively, for the four experiments.

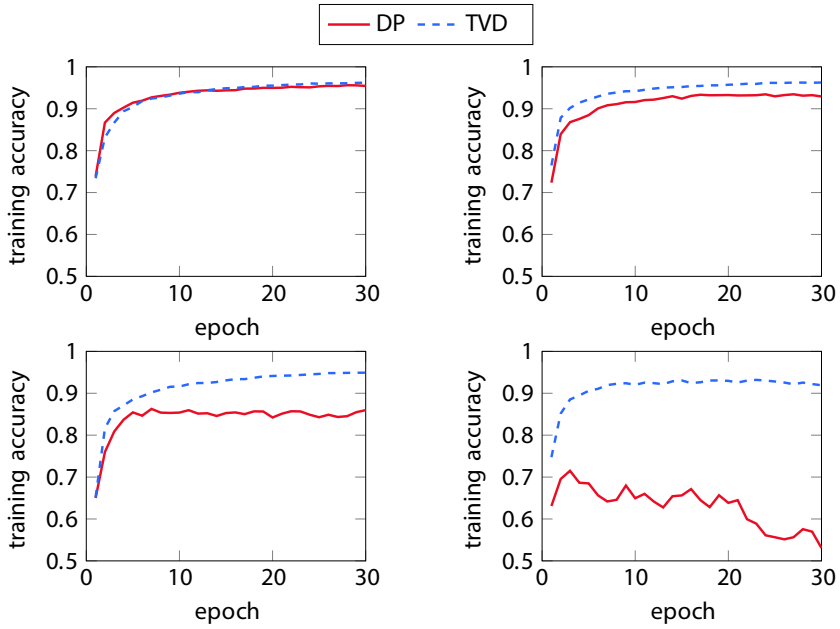

Fig. 4. Training accuracy on the MNIST dataset. The four experiments require the accuracy of the MIAs to be no more than $0.8, 0.7, 0.6$ and 0.55, respectively. In all experiments, the networks are trained using batch size 300, learning rate 0.05 and clipping bound 1. The noise levels of DP and TVD privacy for training the convolutional neural networks are set at $(1.4, 0.7)$, $(2.0, 0.8)$, $(3.7, 1.4)$ and $(7, 2.4)$, respectively, for the four experiments.

The two experiments above validate our theoretical analysis in Section 4.1 that TVD privacy is a suitable measure of the accuracy of adversary inference. Therefore, when we use the accuracy of MIAs to measure the privacy risk of machine learning, TVD privacy always provides a better utility-privacy trade-off.

5.3. Membership Advantage, ROC Curve and Positive Predictive Value

We also analyze other indicators commonly used in MIAs. According to the prior probability, we classify the indicators as balanced and skewed. For these indicators, TVD privacy is still a more accurate metric than DP in most conditions. First, we

consider two indicators under the balanced prior: membership advantage and the ROC curve. Membership advantage and the ROC curve (AUC) are usually used to measure the privacy of machine learning when an adversary has the same prior probability on target samples [9,10,15].

Yeom et al. [9] defined the membership advantage (*Adv*) as the difference between the true positive rate and the false positive rate of MIA. Jayaraman [10] provided a tight membership advantage bound for DP.

Theorem 14 (*Tight Membership Advantage Bounds for DP [10]*). Let $\mathscr{M}$ be an (ε, δ)-DP algorithm. For any randomly chosen record z and fixed false positive rate γ, the membership advantage of a membership inference adversary $\mathscr{A}$ is bounded by:

$$Adv_{\mathscr{A}}(\gamma) \leqslant 1 - f_{\varepsilon,\delta}(\gamma) - \gamma,$$

where $f_{\varepsilon,\delta}(\gamma) = max\{0,\ 1-\delta-e^{\varepsilon}\gamma,\ e^{-\varepsilon}(1-\delta-\gamma)\}$.

Similarly, we can use TVD privacy to limit the membership advantage.

Theorem 15 (*Membership Advantage Bound for TVD Privacy*). Let $\mathscr{M}$ be an α-TVD private algorithm. For any randomly chosen record z and fixed false positive rate γ, the membership advantage of a membership inference adversary $\mathscr{A}$ is bounded by:

$$Adv_{\mathscr{A}}(\gamma) \leqslant min\{1-\gamma,\ \alpha\}.$$

An equivalent bound on membership advantage from TVD privcay appears in [14].

By Theorem 14 and Theorem 15, we plot the curves of DP and TVD privacy in Fig. 5. Since DP is a privacy metric in the worst case, it provides a more accurate measure when the false positive rate is approaching the limit., and TVD privacy provides a more accurate measure of the membership advantage in the middle interval.

In Fig. 6, we plot the ROC curves corresponding to DP and TVD privacy. Fig. 6 shows a similar result to the membership advantage. DP provides a more precise guarantee when the false positive rate is approaching the limit. TVD privacy gives a more accurate guarantee when the difference between the true positive and false positive rates is large.

We also compare the performance of TVD privacy and DP under a skewed prior. Jayaraman et al. [10] considered a more general setup where only a small fraction of the target samples are training samples. They introduced a metric called the positive predictive value (PPV), which is the ratio of correct predictions among all positive predictions. Theorem 16 provides a method to calculate the PPV bound under DP.

Theorem 16 (*Bound of PPV under DP [10]*). Let $\mathscr{M}$ be an (ε, δ)-DP algorithm. For any randomly chosen record z and fixed false positive rate γ, the positive predictive value of a membership inference adversary $\mathscr{A}$ is bounded by

$$PPV_{\mathscr{A}}(\gamma, \rho) \leqslant \frac{1 - f_{\varepsilon,\delta}(\gamma)}{1 - f_{\varepsilon,\delta}(\gamma) + \rho\gamma},$$

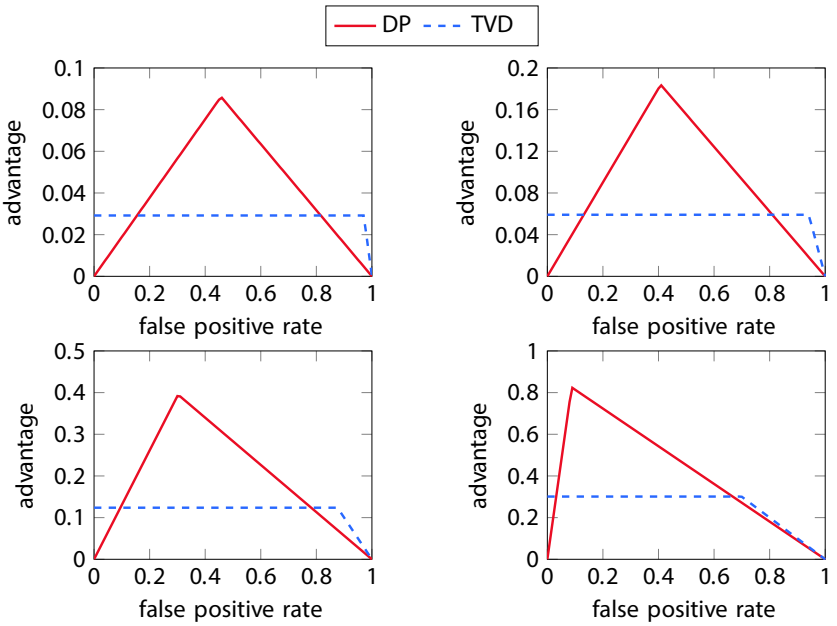

Fig. 5. membership advantage curve of DP and TVD privacy. Each subfigure corresponds to a different σ value at $8, 4, 2$ and 1.

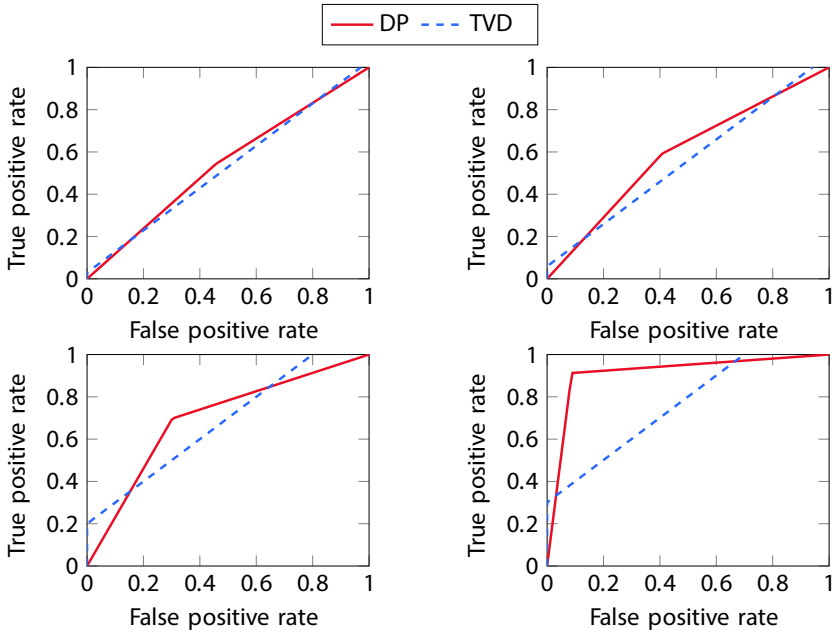

Fig. 6. ROC curve for DP and TVD privacy. Each subfigure corresponds to a different σ value at $8, 4, 2$ and 1.

where $f_{\varepsilon,\delta}(\gamma) = max\{0,\ 1-\delta-e^{\varepsilon}\gamma,\ e^{-\varepsilon}(1-\delta-\gamma)\}$, $\rho = (1-p)/p$, and p is the ratio of the training sample to all test samples.

Similarly, we can also use TVD privacy to bound PPV.

Theorem 17 (*Bound of PPV under TVD Privacy*). Let $\mathscr{M}$ be an α- TVD private algorithm. For any randomly chosen record z and fixed false positive rate γ, the positive predictive value of a membership inference adversary $\mathscr{A}$ is bounded by

$$PPV_{\mathscr{A}}(\gamma,\rho) \leqslant \frac{\gamma+\alpha}{\gamma+\alpha+\rho\gamma}.$$

By Theorem 16 and Theorem 17, we plot the curves of DP and TVD privacy as shown in Fig. 7. In Fig. 7, the Gaussian mechanism adds noise with $\sigma = 2$, and the proportions of training samples are $0.5, 0.2, 0.1$, and 0.01. The conclusion is the same as those of the previous two metrics. DP provides a more accurate measure when the false positive rate is small or large, and TVD privacy offers a more precise estimate in the middle interval. When the proportion of testing samples is low, the upper bounds for DP and TVD privacy are almost identical.

Implementation. From the analysis in this section, TVD privacy has a clear advantage in measuring the accuracy of MIAs, and TVD privacy and DP are complementary in measuring the upper bounds for other indicators of MIAs. Therefore, we recommend choosing TVD privacy for measuring MIA accuracy and combining TVD privacy with DP for other metrics. We provide source code for calculating TVD privacy in private machine learning. Like DP, TVD privacy can be used as a machine learning component. It can give a TVD privacy guarantee for each epoch. We also provide the source code for calculating four indicators in this section with TVD privacy and DP. With these tools, users can easily calculate the theoretical upper bound of each indicator and find a better utility-privacy trade-off. The source code is available from https://github.com/con-fide/TVDprivacy.

6. Discussion

6.1. Optimal composition analysis

In this section, we study the optimal composition of TVD privacy. This study is complementary to Kairouz's work [34]. The Laplace mechanism is taken as an example. The Laplace mechanism with $b = 1$ and a query function with sensitivity $\Delta f = 1$ are considered. We plot the logarithmic ratio of the privacy guarantee of the Laplace mechanism in the worst case in Fig. 8. For most of the interval, the Laplace mechanism ensures that the logarithmic ratio equals ε or $-\varepsilon$. With this feature, the Laplace mechanism is often regarded as the best. However, we focus on values within the $(0, 1)$ interval. Within this interval, the Laplace mechanism provides better privacy guarantees than ε. However, there are no indicators to describe privacy in $(0, 1)$. In other words, this part of privacy protection is wasted in DP.

This observation raises our interest: what will happen if privacy is not wasted? In [34], Kairouz et al. argued that a mechanism without privacy waste has the optimal utility. Furthermore, in [26], Kairouz et al. proved that this mechanism pro-

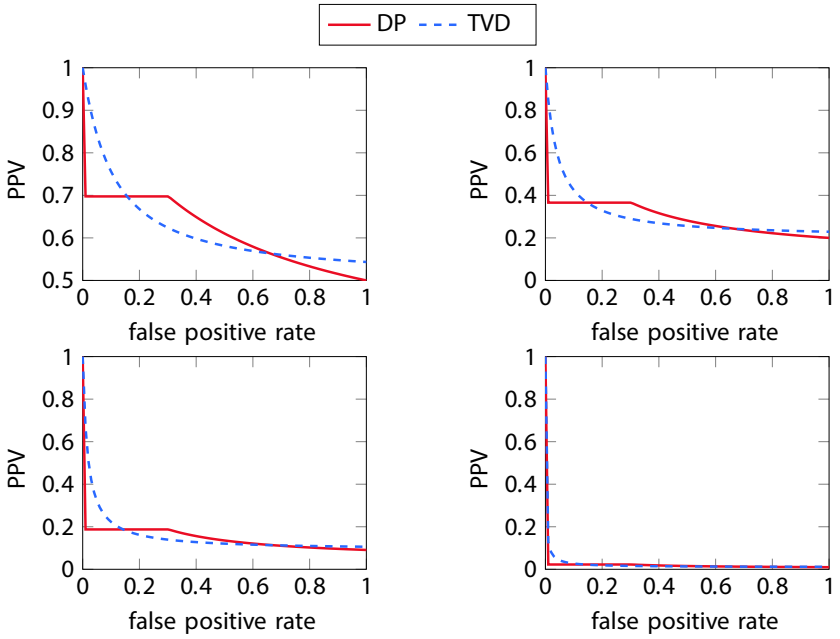

Fig. 7. PPV curve of DP and TVD privacy. Each subfigure corresponds to a different σ value at $8, 4, 2$ and 1.

vides the upper bound of the composition theorem under DP. Our findings further complement their study. When mechanisms satisfy the same TVD privacy properties, the $(\varepsilon, 0)$-DP mechanism without privacy waste has the best composition property. This finding implies that the randomized response mechanism [35] and geometric mechanism [36] have optimal composition properties. First, we review Kairouz's [26] definition of privacy regions under hypothesis testing:

$$R(\varepsilon, \delta) = \{(P_{FN}, P_{FP}) | P_{FP} + e^{\varepsilon} P_{FN} > 1 - \delta, and\ e^{\varepsilon} P_{FP} + P_{FN} > 1 - \delta\},$$

where P_{FN} denotes the false negative rate and P_{FP} denotes the false positive rate.

Similarly, Kairouz defines the *privacy region* of a randomized mechanism $\mathscr{M}$ with respect to two neighboring datasets X_0 and X_1 as

$$R(\mathscr{M}, X_0, X_1) \equiv conv(\{P_{FN}(X_0, X_1, \mathscr{M}, S), P_{FN}(X_0, X_1, \mathscr{M}, S) |\ for\ all\ S \subset \mathbb{X}\}),$$

where S denotes the rejection region.

Kairouz's study implies that we need to find the mechanism $\mathscr{M}$ that minimizes the privacy region $R(\mathscr{M}, X_0, X_1)$ during hypothesis testing. Mechanism $\mathscr{M}$ has an optimal composition property as long as we prove that $R(\mathscr{M}, X_0, X_1) \subset R(\mathscr{M}\prime, X_0, X_1)$ for any other mechanism. We formally present the optimal composition theorem under TVD privacy.

Theorem 18. For any $\alpha \in [0, 1]$, the α-TVD private mechanism $\tilde{\mathscr{M}}_i$ satisfies that $(\varepsilon, 0)$-DP for $\varepsilon = \log \frac{1+\alpha}{1-\alpha}$ has the optimal composition property. For all $i \in \{0, 1, \cdots, \lfloor k/2 \rfloor\}$, the combination mechanism $\tilde{\mathscr{M}} = (\tilde{\mathscr{M}}_1, \cdots, \tilde{\mathscr{M}}_k)$ satisfies $((k - 2i)\varepsilon, \delta)$-DP where

$$\delta_i = \frac{\sum_{l=0}^{i-1} \binom{k}{l} \left(e^{(k-l)\varepsilon} - e^{(k-2i+l)\varepsilon}\right)}{(1+\varepsilon)^k},$$

and satisfies $D_{TV}(B(k, (1+\alpha)/2), B(k, (1-\alpha)/2))$-TVD privacy where $B(n, p)$ denotes the binomial distribution with repetitions n and success probability p.

Proof 10. The proof in this paper is inspired by the supplementary material in [26]. Suppose a mechanism satisfies α-TVD privacy and $(\log \frac{1+\alpha}{1-\alpha}, 0)$-DP. We propose the following mechanism $\tilde{\mathscr{M}}_i$ at the ith step in the composition. For two datasets X_0 and X_1 that differ in only one record, let $\varepsilon = \frac{1+\alpha}{1-\alpha}$, and mechanism $\tilde{\mathscr{M}}_i$ outputs the following:

$$Pr[\tilde{\mathscr{M}}_i(X_0) = y] = \widetilde{P}_0(y) = \begin{cases} \frac{e^{\varepsilon}}{1+\varepsilon} & , for\ y = 0; \\ \frac{1}{1+\varepsilon} & , for\ y = 1; \end{cases}$$

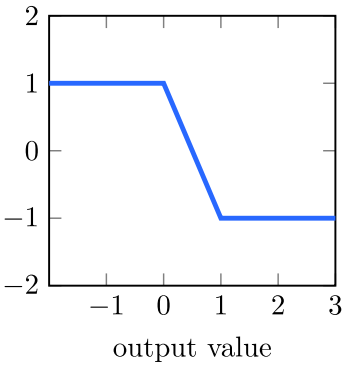

Fig. 8. Actual privacy guarantees for Laplace mechanism. The horizontal axis represents the output domain of the Laplace distribution, and the vertical axis represents the logarithm of the ratio of the probability densities for the two distributions with inputs of 0 and 1..

$$Pr\left[\tilde{\mathscr{M}}_i(X_1)=y\right]=\widetilde{P}_1(y)=\begin{cases}\frac{1}{1+\varepsilon} & ,\text{for } y=0;\\ \frac{e^{\varepsilon}}{1+\varepsilon} & ,\text{for } y=1.\end{cases}$$

We can explicitly compute the composition of mechanism $\tilde{\mathscr{M}}=(\tilde{\mathscr{M}}_1,\tilde{\mathscr{M}}_2,\cdots,\tilde{\mathscr{M}}_k)$. For all $i=\{0,1,\cdots,\lfloor k/2\rfloor\}$, mechanism $\tilde{\mathscr{M}}$ satisfies $((k-2i)\varepsilon,\delta_i)$- DP, where

$$\delta_i=\frac{\sum_{l=0}^{i-1}\frac{k}{l}\left(e^{(k-l)\varepsilon}-e^{(k-2i+l)\varepsilon}\right)}{(1+\varepsilon)^k}. \tag{6}$$

With the help of Kairouz's work, we need only to prove that $R(\tilde{\mathscr{M}}_i,X_0,X_1)\subset R(\mathscr{M}_i,X_0,X_1)$ for any mechanism $\mathscr{M}_i$ that satisfies α-TVD privacy. Let P_0,P_1 denote the probability density function of the outputs $\mathscr{M}(X_0)$ and $\mathscr{M}(X_1)$. $R(\mathscr{M}_i,X_0,X_1)$ is axially symmetric with the function $y=x$. Any mechanism $\mathscr{M}_i$ intersects $\tilde{\mathscr{M}}_i$ at point $(\frac{1-\alpha}{2},\frac{1-\alpha}{2})$. Since $R(\mathscr{M}_i,X_0,X_1)$ is a convex set, the privacy region $R(\mathscr{M}_i,X_0,X_1)$ of any mechanism $\mathscr{M}_i$ contains $R(\tilde{\mathscr{M}}_i,X_0,X_1)$.

Inspired by Theorem 18, we analyze the Laplace and Gaussian mechanisms. Suppose the same single-dimensional numerical query is protected with the Laplace and Gaussian mechanisms. Which mechanism has a better composition property? We plot graphical representations of the Laplace and Gaussian mechanisms satisfying the same α-TVD privacy in Fig. 9. The four plots correspond to $\alpha=0.05,0.1,0.2,0.4$. When both mechanisms satisfy the same TVD privacy property, and the Laplace mechanism satisfies DP with $\varepsilon<2.5$, the privacy region of the Gaussian mechanism always contains the privacy region of the Laplace mechanism. In other words, satisfying the same TVD privacy, the composition property of the Gaussian mechanism is always worse than that of the Laplace mechanism. We compare the privacy regions of the mechanism with the optimal composition property, the Laplace mechanism, and the Gaussian mechanism in Fig. 10. All three mechanisms satisfy 0.1-TVD privacy, and the four plots correspond to the privacy region at the composition numbers $1,10,50,100$. The privacy region of the Gaussian mechanism is significantly larger than that of the Laplace mechanism, and the privacy region of the Laplace mechanism is close to optimal.

6.2. Additional constraints on TVD privacy

Compared to DP, the relaxation of TVD privacy is sometimes unsatisfactory. For some mechanisms, although the parameter of TVD privacy is small, each output may directly reveal the real information of some records. We explain this issue with the following well-known example.

Example 1. Consider an average query function; every record x_i is unique in the dataset D. A randomized mechanism $\mathscr{M}$ randomly selects one of n records as the average. For the neighboring datasets D and $D\prime$, the TVD value between $\mathscr{M}(D)$ and $\mathscr{M}(D\prime)$ is $\alpha=1/n$. When n is large, the randomized mechanism $\mathscr{M}$ seems sufficiently private regarding TVD privacy. However, each output of the randomized mechanism $\mathscr{M}$ leaks a record.

Example 1 implies that TVD privacy may lead to some unpalatable mechanism, which is also the intuition for the privacy parameter δ setting of DP. Mechanism $\mathscr{M}$ in Example 1 is $(0,1/n)$-DP and may leak one record per analysis. Therefore, it is generally accepted that δ in DP should be much less than $1/n$.[5]

[5] This setting ensures that for any mechanism, the probability of leaking one piece of data out of n pieces of data is minimal. McSherry has a different view on the δ setting. He believes that δ should be related to the size of the domain of the input [37]. In either view, the core is to make the probability of privacy loss greater than ε as small as possible.

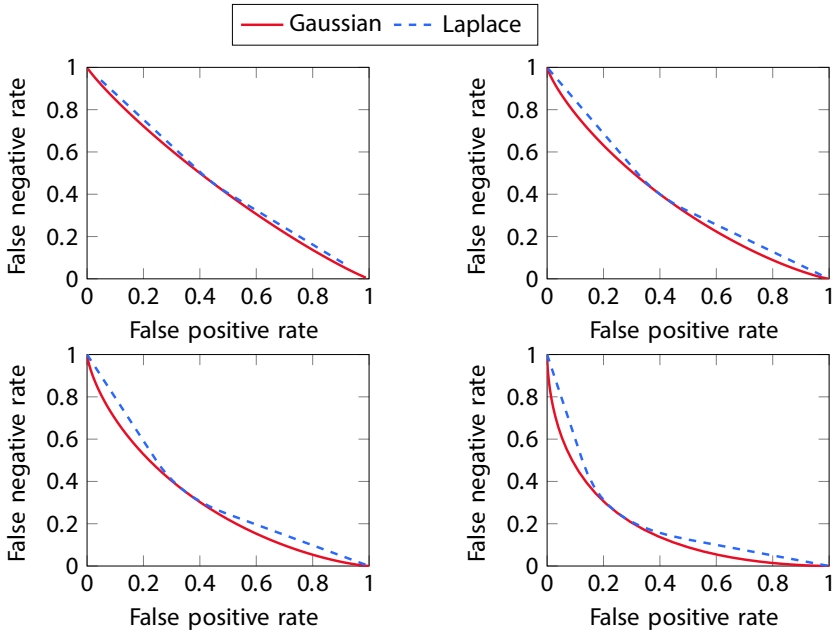

Fig. 9. Comparison of Laplace mechanism and Gaussian mechanism corresponds to a different α value at $0.05, 0.1, 0.2$ and 0.4.

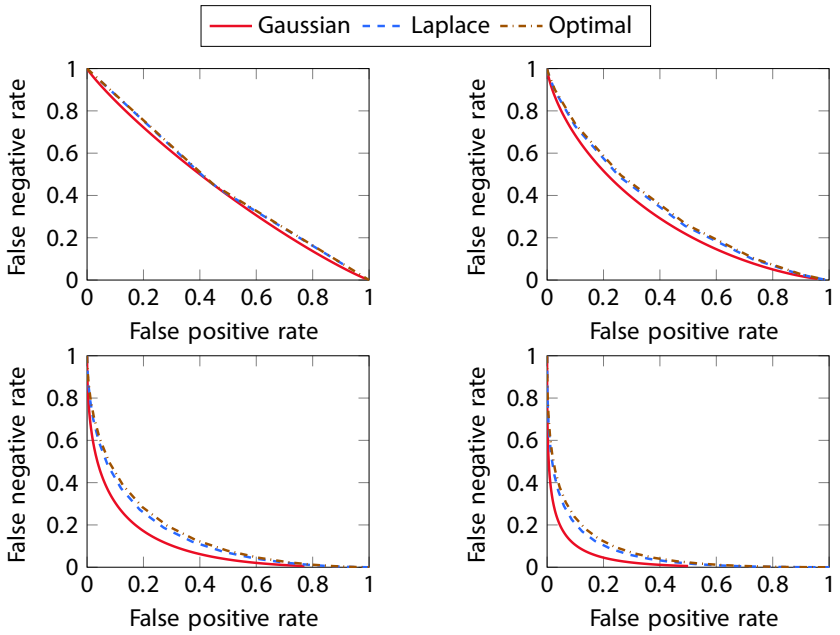

Fig. 10. Composition property comparison of Laplace,Gaussian and Optimal mechanism corresponds to a different composition numbers value at $1, 10, 50$ and 100.

Barber and Duchi [38] suggested dealing with the above problem for TVD privacy. They found that TVD privacy did not provide sufficient protection against privacy risk and suggested imposing more robust divergence requirements to the output distribution. In this paper, we do not want to shift the privacy definition to a more complex form such as KL privacy [38] by adding a divergence constraint. We believe a more complex definition can be more challenging to understand and less convenient. In contrast, we prefer to use the divergence as a complementary constraint on TVD privacy. When the KL diver-

gence is finite, or the maximum divergence is limited with high probability (i.e., (ε, δ)-DP), the mechanism can be analyzed with TVD privacy and does not have the privacy risk in Example 1.

7. Related Work

TVD has been discussed in various forms in the academic community. In [39], Dwork et al. discussed TVD together with DP and explained the advantage of DP instead of TVD. They proposed that TVD may lead to the emergence of certain unpleasant privacy mechanisms. These mechanisms can compromise the privacy of a small number of people while ensuring the majority's privacy.

Ganta et al. [40] proposed a definition called semantic privacy in their analysis of the semantics of DP. TVD privacy is equivalent to semantic privacy with uninformed prior-knowledge assumption. Barber and Duchi [38] formally defined TVD privacy for the first time and compared the performance of multiple privacy definitions. They argued that TVD privacy might not be fully satisfactory. One way to address this issue is to impose a stronger divergence requirement on TVD privacy, such as KL divergence privacy. They found that TVD privacy is weaker than DP but can provide more accurate estimates for privacy risks.

Kairouz et al. [34] used TVD as a utility metric in the analysis of the optimality of DP mechanisms. They demonstrated that staircase mechanisms, including the randomized response, maximize TVD with the same DP guarantee. Rassouli and Gündüz [41] proposed a variant of TVD called average TVD. They showed that the average TVD is consistent with the intuitive notion of a privacy measure and can be used as the privacy measure when solving the optimal utility-privacy trade-off problem by a standard linear procedure.

Shokri et al. [1] proposed MIAs to reveal that machine learning models may leak individual records. Some subsequent works used TVD to measure MIAs. Kulynych et al. [14] used TVD to measure the membership advantage of inference attacks in the worst case. Lin et al. [15] analyzed the privacy of GAN-generated samples and used TVD to measure the robustness of the GAN samples to MIAs.

8. Conclusion

We have provide a systematic theoretical analysis of TVD privacy and analyze the ability of TVD privacy to estimate privacy risks with the example of MIAs. Our work demonstrate that TVD privacy is a helpful tool in estimating privacy risks and has the potential to be widely used as a general privacy definition.

CRediT authorship contribution statement

Jingyu Jia: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing. **Chang Tan:** Data curation. **Zhewei Liu:** Investigation. **Xinhao Li:** Investigation. **Zheli Liu:** Conceptualization. **Siyi Lv:** Conceptualization, Writing - review & editing. **Changyu Dong:** Conceptualization.

Data availability

I have shared the link to my code in the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Controversy over privacy targets

A.1. Motivating Example

To illustrate the controversy of privacy targets, we describe a contagious disease example proposed by Kifer [18], with a causal graph proposed by Michael [17]. The example of the data collection and analysis process on the infectious disease is shown in Fig. A.11. There is an infectious disease that will infect the entire family. In total, n users, including 10 from Bob's family, participate in collecting contagious disease information. A randomized response mechanism collects each data point. Each arrow in the diagram represents a relation between the two ends of the arrow. These relations could be some known information, an unknown correlation, or the direction of the data flow. Arrow (1) represents the family members' infection statuses (sick or not). Arrow (2) represents a group of relations between Bob's family members and society that are not given. Arrows (3), (4), and (5) represent that each individual status in statuses set S corresponds to a raw data point in the database D. Arrows (6), (7), and (8) represent the DP algorithm that generates an output O with dataset D as input. We assume that the

adversary aims to infer the user's sensitive information. Arrow (9) represents the adversary inferring the user's state s by combining his background knowledge θ with the output O.

In this paper, we do not discuss whether DP implicitly assumes data independence, and we are sure there is no such assumption for DP. In this example, we describe the controversy over privacy targets by the maximum successful probability of the adversary's guesses. When Bob participates in the data collection, he receives an ε-DP guarantee. Taking $\varepsilon = 0.5$ as an example, Bob uploads the true value with probability $e^{0.5}/(e^{0.5}+1)$ and uploads the opposite value with probability $1/(e^{0.5}+1)$. By observing Bob's output after a randomized response algorithm, the probability that an adversary without any auxiliary information guesses Bob's message correctly is at most close to 0.63. For an adversary who knows the characteristics of the infectious disease and knows Bob's family's participation in this data collection task, the highest probability of success in guessing Bob's disease is nearly 0.78. The controversy previously centered on whether DP should be criticized for the latter adversary having a successful attack capability of 0.78.

DP believes people should only require privacy protection for the parts they can control. In this example, Bob can neither control the possible auxiliary information of the adversary nor control the family's participation in the data collection task. Therefore, privacy protection for Bob's participation in the data collection task is sufficient. The opposite view believes that "almost no sensitive information about any user should be leaked due to a user answering the query" [22]. In this example, the attributes of Bob's family are directly related to Bob's attributes. Therefore, privacy protection for Bob should consider the impact of Bob's family participation on Bob. We do not judge which of these two views is correct, but we explain the privacy theories behind these two views and analyze how they influence privacy definitions.

A.2. Two Privacy Theories

Privacy is always a broad and ambiguous concept. Many scholars have complained about the difficulty and ambiguity of privacy. Arthur Miller [42] stated that privacy is "difficult to define because it is exasperatingly vague and evanescent." Himma and Tavani [43] detailed discussions on the ethical aspects of information and computers. They argued that "most analyses of issues affecting informational privacy have invoked variations of the restricted access and the control theories". In the previous example, the view that only the participation of Bob needs to be protected adheres to the control theory. The view that Bob's family's data should also be considered is a privacy view from the limited(restricted) access theory. Li et al. [25] noted the conflict between these two privacy theories. In his book, he explains DP's difficulties and ethical challenges. In this paper, we briefly introduce the limited access theorem and control theory.

A.3. Limited Access to Personal Information

Limited access to personal information is one of the most important privacy theories. According to Gavison [44], "privacy is a limitation of others' access to an individual". Similarly, Bok [45] views "privacy as the condition of being protected from unwanted access by others, including physical access, personal information, or attention". In this theory, when a person exists completely independent of society, they acquire absolute privacy. Obviously, people can hardly obtain absolute privacy or complete loss of privacy in society. Therefore, the theory of limited access to personal information is concerned with the "loss of privacy". Tivani and Moor [46] consider that "privacy is fundamentally about protection from intrusion and information gathering by others.".

Limited access theory is an intuitive perception of privacy. If a person does not want a third party to have access to their information, but the third party learns about their information somehow, then their privacy has been violated. According to Himma and Tavani [43], "Arguably, one of the insights of the restricted access theory is in recognizing the importance of zones and contexts that need to be established to achieve informational privacy". In other words, if we wish to guarantee privacy under the limited access theory, we need to determine the scope of privacy and the various types of information that are relevant to the privacy we wish to guarantee.

The main criticism of the theory of limited access to personal information focuses on the lack of a clear distinction between private information and public information [43,47]. Therefore, what level of contact would constitute a privacy violation is difficult to discern. In the example, we assume that the adversary has more auxiliary information. Suppose Bob's property is related to the total number of people with the disease. Bob must be sick if more than a quarter of the population is sick. However, for utility, we offer few reasonable guarantees of privacy in such cases.

A.4. Control over Personal Information

Control over personal information is another important privacy theory. The originator of this theory is Alan Westin [48], who is also the originator of individualistic privacy. Alan Westin stated "Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others". Alan Westin observed that "individuals have needs for disclosure and companionship, which are every bit as important as their needs for privacy". This observation implies that humans need to balance information disclosure with personal privacy. This concept is very common in real life. A person may want to spend time with their family or friends at one time, and at another time, they may want to be left alone. Alan Westin believes that too much or too little privacy can destroy the balance and seriously affect people's lives. When people desire control beyond their power (too much) or when environmental

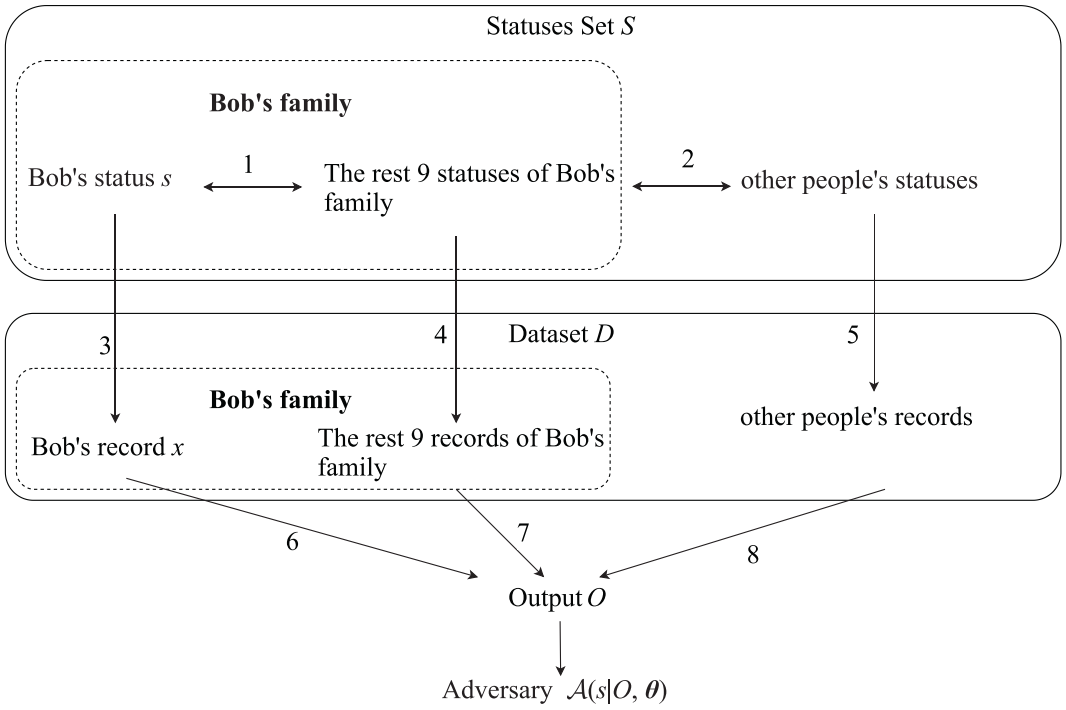

Fig. A.11. Data collection and analysis process on the infectious disease.

factors severely compromise their control over personal information (too little), it is difficult to maintain a privacy balance. Since everyone needs to adjust between solitude and companionship constantly, Alan Westin believes "the individual to decide for himself, with only extraordinary exceptions in the interests of society, when and on what terms his acts should be revealed to the general public" is the core of the right of individual privacy.

Control over personal information theory is shared by some scholars [42,49–51]. Charles Fried makes a clear distinction between the theories of limited access and control over personal information. He argues that "It is not true, for instance, that the less that is known about us, the more privacy we have. Privacy is not simply an absence of information about us in the minds of others; rather it is the control we have over information about ourselves".

As Solove [47] stated, "The control-over-information conception can be viewed as a subset of the limited-access conception". Due to the narrow interpretation of privacy, control over personal information theory has been criticized by many scholars. Tivani and Moor believe that personal control is very limited. Control over personal information theory excludes much of what should be comprehended as privacy. They suggested that even if people lose control of their sensitive information, it should be protected as private [46]. Paul Schwartz [52] is worried that people under control over personal information theory are at a disadvantage to companies in negotiations over the use of personal data. He also realized that "online industry also seeks to lock in a poor level of privacy through collaborative standard setting".

The conflict between the two theories becomes the focus of conflict in the examples. The lack of a direct connection between DP and limited access theory leads to a series of misunderstandings and misapplications. We hope that the work in this section will help the reader to understand privacy better. We also suggest that authors first identify reasonable privacy targets when designing privacy mechanisms or attack algorithms because identifying a reasonable privacy target is the foundation of privacy work.

Appendix B. Proof of Theorem 5.8

For any mechanism $\mathcal{M}$ that satisfies α-TVD privacy, if the input is sampled from the dataset X with a probability q, denoted as $S_q(X)$, then $\mathcal{M} \circ S_q$ satisfies $(q\alpha)$-TVD privacy.

Proof 11. This proof is inspired by Appendix A.1 of [30]. Without loss of generality, assume there are two neighboring datasets X and X', where X' does not contain the record x_n. T denote the sampled dataset. When T does not contain x_n, we have

$$Pr[S_q(X) = T] = Pr[S_q(X') = T]. \tag{B.1}$$

When T contains x_n, T_{-x_n} denotes the sampled dataset with x_n removed. We have

$$Pr\left[S_q(X) = T_{-x_n}\right] = Pr\left[S_q(X') = T_{-x_n}\right]. \tag{B.2}$$

For any possible output set O, let $Z = Pr\left[\mathscr{M}(S_q(X)\right] \in O)$ and $Z' = Pr\left[\mathscr{M}(S_q(X')\right] \in O)$,we have:

$$Z = \sum_{T\subset X} Pr\left[S_q(X) = T\right] Pr[\mathscr{M}(T) \in O],$$
$$Z' = \sum_{T\subset X'} Pr\left[S_q(X) = T\right] Pr[\mathscr{M}(T) \in O].$$

By combining Eq. B.1 and Eq. B.2, we can further describe Z as follows:

$$\begin{aligned} Z &= \sum_{T\subset X} Pr[S_q(X) = T] Pr[\mathscr{M}(T) \in O] \\ &= \sum_{T\subset X'} (1-q) Pr[S_q(X) = T] Pr[\mathscr{M}(T) \in O] + \sum_{T_{-x_n}\subset X'} q Pr[S_q(X) = T_{-x_n}] Pr[\mathscr{M}(T) \in O]. \end{aligned}$$

Let $Y = \sum_{T_{-x_n}\subset X'} Pr\left[S_q(X) = T_{-x_n}\right] Pr[\mathscr{M}(T) \in O]$; then, we have

$$Y \le \sum_{T_{-x_n}\subset X'} Pr[S_q(X) = T_{-x_n}](Pr[\mathscr{M}(T_{-x_n}) \in O] + \alpha) = \sum_{T\subset X'} Pr[S_q(X) = T](Pr[\mathscr{M}(T) \in O] + \alpha) = Z' + \alpha.$$

Then, we have

$$Z = (1-q)Z' + qY \leqslant (1-q)Z' + q(Z+\alpha) = Z + q\alpha.$$

Similarly, we have $Z + q\alpha \geqslant Z'$. We can also use Theorem 9 to prove Theorem 10. S_q satisfies q-TVD privacy, and $\mathscr{M}$ satisfies α- TVD privacy. By Theorem 9, we directly obtain that $\mathscr{M} \circ S_q(D)$ satisfies $(q\alpha)$- TVD privacy

Appendix C. Differentially Private Stochastic Gradient Descent

Algorithm 1 outlines the basic method for training a model with parameters θ by minimizing the empirical loss function $L(\theta)$. At each step of SGD, we compute the gradient $\Delta_\theta L(\theta; x_i)$ for a random subset of examples, clip the l_2 norm of each gradient, compute the average, add noise to protect privacy, and take a step in the opposite direction of this average noisy gradient.

Algorithm 1.: Differentially private SGD [6]

Input: Examples $\{x_1, \cdots, x_N\}$, loss function $L(\theta) = \frac{1}{N}\sum_i L(\theta, x_i)$.
Parameters: learning rate η_t, noise scale σ, group size L,
gradient norm bound C.

Initialize θ_0 randomly
for $t \in [T]$ **do**
  Take a random sample L_i with sampling probability L/N
  Compute gradient
  For each $i \in L_t$, compute $g_t(x_i) \leftarrow \Delta_{\theta_t} L(\theta_t, x_i)$
  Clip gradient
  $\overline{g}_t(x_i) \leftarrow g_t(x_i)/max\left(1, \frac{\|g_t(x_i)\|_2}{C}\right)$
  Add noise
  $\tilde{g}_t \leftarrow \frac{1}{L}\left(\sum_i \overline{g}_t(x_i) + N(0, \sigma^2 C^2 I)\right)$
  Descent
  $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{g}_t$
end
Output: θ_T

References

[1] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE Symposium on Security and Privacy, IEEE Computer Society, San Jose, CA, USA, 2017, pp. 3–18.
[2] Z. Li, Y. Zhang, Membership leakage in label-only exposures, in: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, 2021, pp. 880–895.
[3] M. Fredrikson, S. Jha, T. Ristenpart, Model inversion attacks that exploit confidence information and basic countermeasures, in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 2015, pp. 1322–1333.
[4] M. Fredrikson, E. Lantz, S. Jha, S.M. Lin, D. Page, T. Ristenpart, Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing, in: Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, 2014, pp. 17–32.
[5] J. Gao, B. Hou, X. Guo, Z. Liu, Y. Zhang, K. Chen, J. Li, Secure aggregation is insecure: Category inference attack on federated learning, IEEE Transactions on Dependable and Secure Computing (2021).
[6] M. Abadi, A. Chu, I.J. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 2016, pp. 308–318.
[7] I. Mironov, K. Talwar, L. Zhang, Rényi differential privacy of the sampled gaussian mechanism, CoRR abs/1908.10530 (2019).
[8] M. Nasr, S. Song, A. Thakurta, N. Papernot, N. Carlini, Adversary instantiation: Lower bounds for differentially private machine learning, in: 42nd IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 2021, pp. 866–882.
[9] S. Yeom, I. Giacomelli, M. Fredrikson, S. Jha, Privacy risk in machine learning: Analyzing the connection to overfitting, in: in: 31st IEEE Computer Security Foundations Symposium, IEEE Computer Society, Oxford, United Kingdom, 2018, pp. 268–282.
[10] B. Jayaraman, L. Wang, K. Knipmeyer, Q. Gu, D. Evans, Revisiting membership inference under realistic assumptions, Proceedings on Privacy Enhancing Technologies 2021 (2) (2021) 348–368.
[11] B. Jayaraman, D. Evans, Evaluating differentially private machine learning in practice, in: 28th USENIX Security Symposium, Santa Clara, CA, USA, 2019, pp. 1895–1912.
[12] M.A. Rahman, T. Rahman, R. Laganière, N. Mohammed, Membership inference attack against differentially private deep learning model, Transactions on Data Privacy 11 (1) (2018) 61–79.
[13] T. Humphries, M. Rafuse, L. Tulloch, S. Oya, I. Goldberg, F. Kerschbaum, Differentially private learning does not bound membership inference, CoRR abs/2010.12112 (2020).
[14] B. Kulynych, M. Yaghini, G. Cherubin, M. Veale, C. Troncoso, Disparate vulnerability to membership inference attacks, Proceedings on Privacy Enhancing Technologies 2022 (1) (2022) 460–480.
[15] Z. Lin, V. Sekar, G. Fanti, On the privacy properties of gan-generated samples, in: The 24th International Conference on Artificial Intelligence and Statistics, Vol. 130 of Proceedings of Machine Learning Research, PMLR, 2021, pp. 1522–1530.
[16] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Foundations and Trends in Theoretical Computer Science 9 (3–4) (2014) 211–407.
[17] M.C. Tschantz, S. Sen, A. Datta, Sok: Differential privacy as a causal property, in: 2020 IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 2020, pp. 354–371.
[18] D. Kifer, A. Machanavajjhala, No free lunch in data privacy, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, Athens, Greece, 2011, pp. 193–204.
[19] F. McSherry, Lunchtime for data privacy, https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-16.md (2016).
[20] D. Kifer, A. Machanavajjhala, Pufferfish: A framework for mathematical privacy definitions, ACM Transactions on Database Systems 39 (1) (2014) 3:1–3:36.
[21] B. Yang, I. Sato, H. Nakagawa, Bayesian differential privacy on correlated data, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, 2015, pp. 747–762.
[22] J. Zhao, J. Zhang, H.V. Poor, Dependent differential privacy for correlated data, in: 2017 IEEE Globecom Workshops, Singapore, 2017, pp. 1–7.
[23] R. Chen, B.C.M. Fung, P.S. Yu, B.C. Desai, Correlated network data publication via differential privacy, The VLDB Journal 23 (4) (2014) 653–676.
[24] P. Rogaway, The moral character of cryptographic work, IACR Cryptology ePrint Archive (2015) 1162.
[25] N. Li, M. Lyu, D. Su, W. Yang, Differential Privacy: From Theory to Practice, Synthesis Lectures on Information Security, Privacy, & Trust, Morgan & Claypool Publishers, 2016.
[26] P. Kairouz, S. Oh, P. Viswanath, The composition theorem for differential privacy, in: Proceedings of the 32nd International Conference on Machine Learning, Vol. 37 of JMLR Workshop and Conference Proceedings, JMLR.org, Lille, France, 2015, pp. 1376–1385.
[27] K. Grining, M. Klonowski, Towards extending noiseless privacy: Dependent data and more practical approach, in: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, Abu Dhabi, United Arab Emirates, 2017, pp. 546–560.
[28] D. Desfontaines, E. Mohammadi, E. Krahmer, D. Basin, Differential privacy with partial knowledge, arXiv preprint arXiv:1905.00650 (2019).
[29] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, S. Song, K. Talwar, A. Thakurta, Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation, CoRR abs/2001.03618 (2020).
[30] N. Li, W.H. Qardaji, D. Su, On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy, in: 7th ACM Symposium on Information, Computer and Communications Security, Seoul, Korea, 2012, pp. 32–33.
[31] B. Balle, J. Bell, A. Gascón, K. Nissim, The privacy blanket of the shuffle model, in: 39th Annual International Cryptology Conference, Vol. 11693 of Lecture Notes in Computer Science, Santa Barbara, CA, USA, 2019, pp. 638–667.
[32] A. Bittau, Ú. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnés, B. Seefeld, Prochlo: Strong privacy for analytics in the crowd, in: Proceedings of the 26th Symposium on Operating Systems Principles, Shanghai, China, 2017, pp. 441–459.
[33] A. Cheu, A.D. Smith, J.R. Ullman, D. Zeber, M. Zhilyaev, Distributed differential privacy via shuffling, in: 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vol. 11476 of Lecture Notes in Computer Science, Darmstadt, Germany, 2019, pp. 375–403.
[34] P. Kairouz, S. Oh, P. Viswanath, Extremal mechanisms for local differential privacy, in: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, Montreal, Quebec, Canada, 2014, pp. 2879–2887.
[35] S.L. Warner, Randomized response: A survey technique for eliminating evasive answer bias, Journal of the American Statistical Association 60 (309) (1965) 63–69.
[36] A. Ghosh, T. Roughgarden, M. Sundararajan, Universally utility-maximizing privacy mechanisms, SIAM Journal on Computing 41 (6) (2012) 1673–1693.
[37] F. McSherry, Two flavors of differential privacy, https://github.com/frankmcsherry/blog/blob/master/posts/2017-02-08.md (2017).
[38] R.F. Barber, J.C. Duchi, Privacy and statistical risk: Formalisms and minimax bounds, arXiv preprint arXiv:1412.4451 (2014).
[39] C. Dwork, F. McSherry, K. Nissim, A.D. Smith, Calibrating noise to sensitivity in private data analysis, Theory of Cryptography, Third Theory of Cryptography Conference of Lecture Notes in Computer Science, Vol. 3876, Springer, New York, NY, USA, 2006, pp. 265–284.
[40] S.R. Ganta, S.P. Kasiviswanathan, A.D. Smith, Composition attacks and auxiliary information in data privacy, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, 2008, pp. 265–273.
[41] B. Rassouli, D. Gündüz, Optimal utility-privacy trade-off with total variation distance as a privacy measure, IEEE Transactions on Information Forensics and Security 15 (2019) 594–603.
[42] A.R. Miller, The Assault On Privacy: Computers, Data Banks, And Dossiers, 1st Edition,,, Ann Arbor/The University Of Michigan Press, 1971.
[43] K.E. Himma, H.T. Tavani, The handbook of information and computer ethics, John Wiley & Sons, 2008.
[44] R. Gavison, Privacy and the limits of law, The Yale Law Journal 89 (3) (1980) 421–471.

[45] S. Bok, Secrets: On the ethics of concealment and revelation, Vintage, 1989.
[46] H.T. Tavani, J.H. Moor, Privacy protection, control of information, and privacy-enhancing technologies, ACM Sigcas Computers and Society 31 (1) (2001) 6–11.
[47] D.J. Solove, Conceptualizing privacy, Calif. L. Rev. 90 (2002) 1087.
[48] A.F. Westin, Privacy and freedom, Washington and Lee Law Review 25 (1) (1968) 166.
[49] A.A. Miller, The assault on privacy., Psychiatric Opinion (1975).
[50] J. Rachels, Why privacy is important, Philosophy & Public Affairs (1975) 323–333.
[51] C. Fried, Privacy, Philosophical dimensions of privacy (1984) 203–222.
[52] P.M. Schwartz, Privacy and democracy in cyberspace, Vanderbilt Law Review 52 (1999) 1607.