Journal of Econometrics 235 (2023) 444–453

Contents lists available at ScienceDirect

Journal of Econometrics

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

Distribution-invariant differential privacy

Xuan Bi [a,*], Xiaotong Shen [b]

[a] Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, MN, United States of America

[b] School of Statistics, University of Minnesota, Minneapolis, MN, United States of America

ARTICLE INFO

Article history:
Received 17 June 2021
Received in revised form 11 May 2022
Accepted 15 May 2022
Available online 18 June 2022

JEL classification:
C10

Keywords:
Privacy protection
Distribution preservation
Data sharing
Data perturbation
Randomized mechanism

ABSTRACT

Differential privacy is becoming one gold standard for protecting the privacy of publicly shared data. It has been widely used in social science, data science, public health, information technology, and the U.S. decennial census. Nevertheless, to guarantee differential privacy, existing methods may unavoidably alter the conclusion of original data analysis, as privatization often changes the sample distribution. This phenomenon is known as the trade-off between privacy protection and statistical accuracy. In this work, we mitigate this trade-off by developing a distribution-invariant privatization (DIP) method to reconcile both high statistical accuracy and strict differential privacy. As a result, any downstream statistical or machine learning task yields essentially the same conclusion as if one used the original data. Numerically, under the same strictness of privacy protection, DIP achieves superior statistical accuracy in in a wide range of simulation studies and real-world benchmarks.

1. Introduction

Data privacy has become increasingly important in many fields in the big data era (Kearns et al., 2016; Cohen and Nissim, 2020), where a massive amount of sensitive information is digitally collected, stored, transferred, and analyzed. To protect data privacy, differential privacy (Dwork, 2006a) has recently drawn great attention. It quantifies the notion of privacy for downstream machine learning tasks (Jordan and Mitchell, 2015) and protects even the most extreme observations. This quantification is useful for publicly released data such as census and survey data, and improves transparency and accessibility in artificial intelligence (Haibe-Kains et al., 2020). It has been adopted in biomedical research (Kaissis et al., 2020; Jobin et al., 2021; Hie et al., 2018; Han et al., 2020), epidemiology (Venkatramanan et al., 2021), sociology (Santos-Lozada et al., 2020), and by many technology companies, such as Google (Erlingsson et al., 2014), Apple (Apple Differential Privacy Team, 2017), Microsoft (Ding et al., 2017), LinkedIn Kenthapadi and Tran (2018), Amazon (Day One Staff, 2018), and Facebook (Nayak, 2020). In 2020, differential privacy is, for the first time, used to protect the confidentiality of individuals in the U.S. decennial census (United States Census Bureau, 2020).

Scientists have developed various differential private methods. These methods not only protect data privacy but also promote data sharing. This allows privatized data from multiple entities to be integrated into a single model to strengthen data analysis without leaking critical information (Vadhan, 2017). This aspect brings huge economic and societal benefits in the big data era. However, one major issue is that data privatization may alter the analysis result of the original data, and it is believed that there is a trade-off between statistical accuracy and differential privacy (Goroff, 2015; Santos-Lozada et al., 2020; Gong and Meng, 2020; Bowen and Liu, 2020). In other words, one needs to sacrifice the accuracy of a

* Corresponding author.
E-mail addresses: xbi@umn.edu (X. Bi), xshen@umn.edu (X. Shen).

https://doi.org/10.1016/j.jeconom.2022.05.004
0304-4076/

downstream analysis for privacy protection. On the contrary, we show that one can reconcile *both* accuracy and privacy, which we achieve by preserving the original data's distribution. We develop a distribution-invariant privatization method (DIP) that achieves private data release. It transforms and perturbs the data and employs a suitable transformation to recover the original distribution. As a result, DIP maintains statistical accuracy while being differentially private at any desired level of protection.

There is a large body of literature on differential privacy. Two major directions emerge from computer science and statistics. The first direction achieves privacy protection by either a privatization mechanism or a privatized sampling method, including the Laplace mechanism (Dwork et al., 2006c; Dwork and Roth, 2014), the exponential mechanism (McSherry and Talwar, 2007), the minimax optimal procedures (Duchi et al., 2018), among others. The second achieves differential privacy via privatization for a category of models or algorithms, such as deep learning (Abadi et al., 2016), boosting (Dwork et al., 2010), stochastic gradient descent (Agarwal et al., 2018), risk minimization (Chaudhuri et al., 2011), random graphs (Karwa and Slavković, 2016), function estimation (Hall et al., 2013), parametric estimation (Avella-Medina, 2021), regression diagnostics (Chen et al., 2016), and top-k selection (Durfee and Rogers, 2019). Moreover, a general statistical framework for differential privacy is introduced (Wasserman and Zhou, 2010), and theoretical properties of various privatization mechanisms are investigated (Kairouz et al., 2017; Vadhan, 2017). Other types of differential privacy as well as the associated statistical properties are also discussed, including relaxed or approximate differential privacy (Dwork et al., 2006b; Abadi et al., 2016; Agarwal et al., 2018; Cai et al., 2019), local differential privacy (Evfimievski et al., 2003; Kasiviswanathan et al., 2011; Ding et al., 2017; Rohde and Steinberger, 2020), random differential privacy (Hall et al., 2012), Renyi differential privacy (Mironov, 2017), and Gaussian differential privacy (Dong et al., 2019).

One main challenge is that existing privatization mechanisms protect data privacy at the expense of altering a sample's distribution. For example, count data may become negative values after adding a noise, and non-trivial or data-dependent post-processing may be needed. As a result, analysis of such privatized data may conclude dramatically different from the analysis of the original data. From a machine learning perspective, developing a distribution-invariant privatization mechanism becomes crucial to maintain statistical accuracy.

This article proposes a DIP mechanism to address the above challenge for essentially *all* types of data, such as those following continuous, discrete, mixed, categorical, empirical, or multivariate distributions. It satisfies differential privacy while approximately preserving the original distribution when it is unknown through a reference distribution, for example, the empirical distribution on a hold-out sample and exactly preserving the original distribution when it is known, c.f., Theorems 1 and 2. It is scalable to massive or high-dimensional data, c.f., Proposition 1. Consequently, any downstream statistical analysis or machine learning tasks will lead to nearly the same conclusion as if the original data were used, which is a unique aspect that existing methods do not share. It allows distributions with unbounded support, which would be otherwise rather difficult if not impossible (Wasserman and Zhou, 2010). Moreover, DIP approximately maintains statistical accuracy even with strict privacy protection in that it does not suffer from the trade-off between accuracy and privacy strictness, as illustrated in our simulation studies in Section 2. These characteristics enable us to perform data analysis without sacrificing statistical accuracy, as in regression, classification, graphical models, clustering, among other statistical and machine learning tasks.

2. Results

2.1. Overview of the proposed method

This subsection presents the main ideas of the proposed DIP method. Differential privacy protects publicly shared information about a dataset by describing its patterns while guarding against disclosing the data information. The reader may consult Definition 1 for a rigorous mathematical definition. In what is to follow, we discuss the privatization of univariate and multivariate samples separately.

DIP's privatization process consists of three steps. First, DIP splits the original sample randomly into two independent subsamples, hold-out and to-be-privatized samples, both are fixed after the split. Second, it estimates an unknown data distribution by, say, the empirical distribution on the hold-out sample, which is referred to as a reference distribution. Third, we privatize the to-be-privatized sample through data perturbation, which (i) satisfies the requirement of differential privacy, and (ii) preserves the reference distribution approximating the original distribution. As a result, DIP is differential private on the to-be-privatized sample while retaining the original distribution asymptotically, c.f., Theorem 2.

For univariate data, the privatization process of a to-be-privatized sample is described in Fig. 1 and Section 4. First, we apply the probability-integral transformation to each data point or its smooth version to yield a uniform distribution. Second, we add a random Laplace noise to perturb and mask the data, where the Laplace noise entails differential privacy (Dwork, 2006a), as opposed to, say, the Gaussian noise. Third, we design a new function transforming the obfuscated data to follow the reference distribution approximating the original data distribution.

For multivariate data, we privatize each variable sequentially, using the univariate method above. To preserve the distribution, we propose to apply the probability chain rule (Schum, 2001), in place of privatizing each variable independently. In other words, we privatize the first variable, and then the second variable using its conditional distribution given the privatized value of the first variable, and so forth; c.f., Section 4.

Finally, we develop a fast and scalable algorithm, c.f., Algorithm 1 and Proposition 1 in Section 4, for practical use, and prove that DIP preserves the data distribution approximately while being differentially private for essentially all types of data, c.f., Theorem 2.

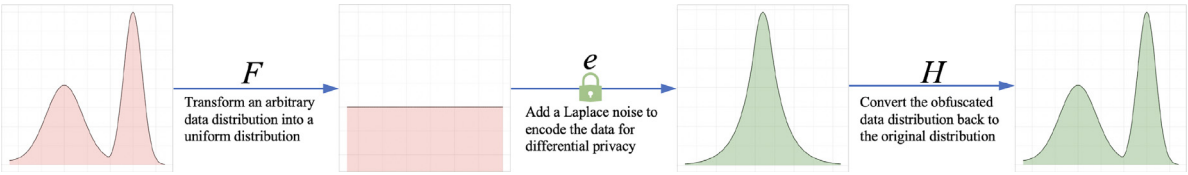

Fig. 1. A flowchart visualizing the proposed DIP method. The left arrow: Data from an arbitrary distribution is transformed into a sample of the uniform distribution via the probability integral transformation function F. The middle arrow: A noise following a centered Laplace distribution is added to the transformed data to ensure differential privacy. The obfuscated data now follow a convoluted distribution, whose explicit form is provided in Appendix S1.1. The right arrow: A carefully designed transformation function H is applied to the obfuscated data to obtain a privatized sample, which follows the original distribution but does not necessarily contain the original data values. Non-private and private distributions are filled in red and green, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2. Simulation studies

This subsection performs simulations to investigate the operating characteristics of DIP and compare it with some top competitors, including the Laplace randomized mechanism (LRM) (Dwork et al., 2006c), the minimax optimal procedure mechanism (OPM) (Duchi et al., 2018), and the exponential mechanism (EXM) (McSherry and Talwar, 2007). For a fair comparison, all competing methods, including DIP, use the same original dataset and the same privacy factor (defined in Definition 1), while assuming the data distribution is unknown. Then the statistical accuracy of a downstream method on privatized data is evaluated. It is worth mentioning that DIP requires sample-splitting, which is at a disadvantage for downstream analysis due to the reduced size of the released sample. Despite the disadvantage, DIP achieves superior statistical accuracy in two simulations and on three real-world benchmarks.

2.2.1. Network structure reconstruction

We first consider the private reconstruction of a network's structure. We generate a random sample from a multivariate normal distribution, in which its precision matrix, or the inverse of the covariance matrix, encodes an undirected graph for the network. Then we estimate the precision matrix based on the privatized random sample.

We examine two types of graph networks, including the chain and exponential decay networks. For the chain network, the precision matrix is a sparse tridiagonal matrix corresponding to the first-order autoregressive structure. For the exponential decay network, elements of the precision matrix are exponentially decayed, and the network is not sparse. We perform 1000 simulations in a high-dimensional setting where the sample size $N = 200$ is smaller than the number of variables $p = 250$. The privacy factor is from 1 to 3. After privatization, we estimate the precision matrix via graphical Lasso (Friedman et al., 2008). The estimated precision matrix is evaluated by the entropy loss (Lin and Perlman, 1985) by comparing it with the true precision matrix. For DIP, we apply (2) with a random split of the original sample where a ratio of 15%, 25%, 35% is used as the hold-out sample. For LRM and OPM, we follow Dwork et al. (2006c) and Duchi et al. (2018) to privatize the entire sample.

As suggested by Table 1, DIP performs well. DIP is insensitive to the value of the privacy factor, which agrees with the theoretical result in Theorem 2. Moreover, neither LRM nor OPM applies because of the requirement of bounded support, as evident by an infinite value of the entropy loss caused by the estimate of the precision matrix being 0. In summary, DIP's distribution preservation property becomes more critical to explore multivariate structures, which explains sizeable differences between DIP and its competitors (LRM and OPM) in simulations.

2.2.2. Linear regression

Consider a linear model in which the sample size is $N = 200$ or 2000 and the design matrix consists of 6 or 30 variables. The regression coefficient is set as a vector of 1's. Each 1/3 of the design matrix's columns follow independently $Normal(0, 10^2)$, $Poisson(5)$, and $Bernoulli(0.5)$, respectively. The privacy factor is 1, 2, 3, or 4. In this example, we estimate the regression coefficient vector based on the privatized release data and measure the estimation accuracy by the Euclidean distance between the estimated and true parameter vectors.

For DIP, we split the data into hold-out and to-be-privatized samples with a splitting ratio of 15%, 25%, and 35%. Then we apply Algorithm 1 to privatize variables of the to-be-privatized sample in a random order to examine DIP's invariant property of the sequential order in Theorem 2. For LRM and OPM, we adopt the univariate case as in the graphical model case. Note that LRM and OPM utilize the additional information — independence among columns of the design matrix, whereas DIP does not. The simulation replicates 1000 times.

As indicated in Table 2, DIP yields the lowest estimation error across all situations with a substantial amount of improvement over LRM and OPM, ranging from 559% to 1071503%, despite that DIP's to-be-privatized sample size is smaller than the size of the entire sample. Importantly, DIP mitigates the trade-off between differential privacy and statistical accuracy, as an increased level of the privacy factor has little impact on the estimation accuracy. Meanwhile, only a small difference is seen between DIP and the oracle non-private counterpart. This is a small price to be paid for estimating an unknown distribution. The small standard error also suggests that the multivariate DIP is invariant against the random order of variable privatization, which agrees with the invariant property for the sequential order in Theorem 2.

Table 1
DIP is the only method that protects the structure of high-dimensional graphical networks. The performance of DIP is measured by the entropy loss with its standard error in the parenthesis. Two types of networks, chain and exponential decay, are evaluated, with a sample size $N = 200$ and the number of variables $p = 250$. A ratio of 15%, 25%, and 35% of the original sample is used as the hold-out sample to estimate the empirical distribution. The privacy factor ε (i.e., the budget of privacy) ranges from 1 to 3. All of the state-of-the-art methods get infinite loss and their results are omitted.

Network Reconstruction		ε		
Setting	Hold-out	1	2	3
Chain Net	15%	101.53 (2.82)	101.51 (2.83)	101.49 (2.82)
$N = 200$	25%	79.21 (2.21)	79.18 (2.19)	79.18 (2.19)
$p = 250$	35%	75.47 (1.75)	75.46 (1.74)	75.45 (1.73)
Exp Decay	15%	42.18 (1.10)	42.17 (1.10)	42.16 (1.09)
$N = 200$	25%	29.32 (1.25)	29.31 (1.25)	29.31 (1.25)
$p = 250$	35%	25.35 (1.41)	25.35 (1.43)	25.33 (1.43)

Table 2
DIP mitigates the trade-off between differential privacy and statistical accuracy. Private linear regression is conducted. Averaged L_2-distance (standard error) between true and estimated regression coefficients are measured across different sample sizes and numbers of variables.

Linear Regression		ε			
Setting	Method	1	2	3	4
$N = 200$	NP	0.31 (0.16)	0.31 (0.15)	0.31 (0.16)	0.31 (0.15)
$p = 6$	DIP (hold 15%)	2.19 (1.15)	2.16 (1.14)	2.14 (1.13)	2.14 (1.14)
	DIP (hold 25%)	1.40 (0.77)	1.40 (0.75)	1.42 (0.79)	1.40 (0.76)
	DIP (hold 35%)	1.13 (0.59)	1.13 (0.61)	1.11 (0.62)	1.12 (0.59)
	LRM	62.80 (43.28)	33.53 (21.43)	24.53 (14.86)	21.44 (11.91)
	OPM	757.16 (558.17)	378.68 (287.99)	241.05 (190.85)	184.72 (136.91)
$N = 2000$	NP	0.09 (0.04)	0.09 (0.04)	0.09 (0.04)	0.09 (0.04)
$p = 6$	DIP (hold 15%)	0.31 (0.15)	0.32 (0.15)	0.31 (0.15)	0.32 (0.15)
	DIP (hold 25%)	0.24 (0.12)	0.24 (0.12)	0.24 (0.12)	0.24 (0.11)
	DIP (hold 35%)	0.21 (0.10)	0.21 (0.10)	0.21 (0.10)	0.20 (0.10)
	LRM	19.55 (13.66)	13.67 (7.94)	12.30 (6.00)	11.83 (4.91)
	OPM	650.15 (489.16)	315.75 (240.77)	219.46 (167.75)	165.44 (125.28)
$N = 200$	NP	0.77 (0.30)	0.74 (0.30)	0.76 (0.32)	0.77 (0.33)
$p = 30$	DIP (hold 15%)	21.07 (10.64)	20.91 (10.68)	21.03 (10.84)	21.05 (10.94)
	DIP (hold 25%)	9.16 (4.41)	9.17 (4.47)	9.28 (4.45)	9.21 (4.46)
	DIP (hold 35%)	6.15 (2.84)	6.24 (2.95)	6.23 (2.95)	6.21 (3.00)
	LRM	489.43 (370.81)	250.98 (182.24)	179.74 (117.89)	138.68 (90.79)
	OPM	6044.5 (4611.4)	2951.1 (2277.7)	2021.9 (1492.6)	1542.4 (1118.0)
$N = 2000$	NP	0.21 (0.09)	0.21 (0.08)	0.21 (0.08)	0.21 (0.08)
$p = 30$	DIP (hold 15%)	1.39 (0.59)	1.39 (0.59)	1.40 (0.60)	1.39 (0.60)
	DIP (hold 25%)	0.88 (0.38)	0.88 (0.39)	0.89 (0.38)	0.88 (0.39)
	DIP (hold 35%)	0.73 (0.32)	0.71 (0.32)	0.72 (0.31)	0.72 (0.33)
	LRM	209.25 (147.50)	116.93 (84.48)	86.97 (57.60)	74.07 (45.71)
	OPM	7822.7 (6055.5)	4046.6 (3073.2)	2595.7 (1944.0)	2052.9 (1522.2)

DIP shows consistent and robust performance against the change of ε. NP represents the non-private benchmark.

2.3. Real data analysis

Next, we analyze three sensitive benchmark datasets to understand the practical implications of distribution-invariant privatization.

2.3.1. The University of California salary data

The first study concerns the University of California system salary data collected in 2010 (University of California, 2014). The dataset contains annual salaries of 252,540 employees, including faculty, researchers, and staff. The average salary of all employees is $39,531.49 with a standard deviation of $53,253.93. The data distribution is highly right-skewed, with the 90% quantile being $95,968.12 and the maximum exceeding two million dollars.

For this study, we estimate the differentially private mean salary. One important aspect is contrasting the privatized mean with the original mean $39,531.49 to understand the impact of privatization on statistical accuracy of estimation. The relative mean difference (i.e., the difference between the private mean and the original mean divided by the original mean) is evaluated. Three privatization mechanisms are compared, including DIP, LRM, and OPM. For DIP, we apply Algorithm 1. For LRM, random Laplace noise is added to the original data before calculating the private mean. For OPM, we follow the private mean estimation function described in Section 3.2.1 of Duchi et al. (2018) to optimize its performance. The above process, including privatization, is repeated 1000 times.

Table 3
DIP shows the best performance in mean estimation, logistic regression, and personalized recommendations across three benchmark datasets. Data description, including the sample size, data type, number of variables, and the desired task, is provided on the top half of the table. The evaluation (with standard errors) of DIP and the state-of-the-art privacy protection methods (LRM, OPM, and EXM), is provided on the bottom half of the table. The evaluation metrics are the relative mean difference, the Kullback–Leibler divergence, and the root mean square error on a random 25% test set for the UC salary data, the bank marketing data, and the MovieLens data, respectively. For each dataset, the same privacy protection strictness is imposed, and a smaller evaluation measure indicates a more accurate result.

	Dataset		
	UC salary	Bank marketing	MovieLens
Size	252,540	30,488	25,000,095
Type	Continuous	Multivariate	Discrete
Dimension	1	10	1
Task	Mean Est.	Logistic reg.	Collaborative filtering
DIP	0.40 (0.31)	0.047 (0.003)	1.03 (4.96×10^{-4})
LRM	13.08 (9.95)	0.311 (0.004)	1.87 (1.03×10^{-3})
OPM	4.69 (3.44)	Infinity	2.61 (8.23×10^{-4})
EXM	N/A	N/A	1.11 (5.52×10^{-4})

N/A means that a result is unavailable.

2.3.2. Portuguese bank marketing campaign data

The second study focuses on a set of marketing campaign data collected from a Portuguese retail bank from 2008 to 2013 (Moro et al., 2014). This campaign intended to sell long-term deposits to potential clients through phone conversations. During a phone call, an agent collected a client's sensitive personal information, including age, employment status, marital status, education level, whether the client has any housing or personal loan, and if the client is in a default status. In addition, the agent collected the client's device type, past contact histories regarding this campaign, and if the client is interested in subscribing to a term deposit (yes/no). The dataset contains a total of 30,488 respondents whose data are complete.

Our goal is to conduct private logistic regression and examine the statistical accuracy change after privatization. For DIP, Algorithm 1 is applied. For LRM, Laplace noise is added to each variable independently. For OPM, we conduct private logistic regression following the private estimation of generalized linear models in Section 5.2.1 of Duchi et al. (2018). The Kullback–Leibler divergence is measured to evaluate the discrepancy between the private and the original logistic regression results. The privatization process repeats 1000 times.

2.3.3. MovieLens data

The third study considers the privacy protection of movie ratings for a private recommender system. Many online platforms disclose de-identified user data for research or commercial purposes. In many situations, however, simply removing user identities is not adequate. For example, as suggested in Narayanan and Shmatikov (2008), political preferences and other sensitive information can still be uncovered based on de-identified movie ratings.

To overcome this difficulty, we consider the MovieLens 25M dataset, which is collected by GroupLens Research (Harper and Konstan, 2015) between 1995 and 2019. It contains 25,000,095 movie ratings, collected from 162,541 users over 59,047 movies. In other words, each user rated 154 out of the 59,047 movies on average, leaving their preference towards the vast majority of movies unknown. Movie ratings are values in $\{0.5, 1, 1.5, \ldots, 5\}$. A typical recommender system intends to predict a user's rating on movies that they have not watched yet, and then recommends movies to each user based on high predicted ratings. The focus here is a prediction as opposed to inference in the previous two studies.

To investigate the effect of data privatization on the prediction accuracy, we randomly split the movie ratings into a 75% training set and a 25% test set, and privatize the training set. For DIP, we apply Algorithm 1. For LPM, Laplace noises are added to the raw ratings, and the noised ratings are rounded to the nearest number in $\{0.5, 1, 1.5, \ldots, 5\}$. For OPM, we follow Section 4.2.3 of Duchi et al. (2018). For EXM, the sampling scheme described in Appendix S2 (in the subsection about the MovieLens data) is applied. Then we train a matrix factorization model (Funk, 2006) based on privatized ratings, which is a prototype collaborative filtering method for movie recommendation. The evaluation metric is the root mean square error on the non-private test set, which is averaged over 50 random partitions of training and test sets.

2.3.4. Analysis results

In all three benchmark examples, the privacy factor is 1. For DIP, we hold out 25% of the sample for estimating the unknown data distribution. As shown in Table 3, DIP delivers desirable statistical accuracy across the three different statistical and machine learning tasks, demonstrating the benefits of distribution-preservation.

For the mean salary estimation in the first example, the discrepancy between the original mean and the private mean of DIP is minimal. Specifically, while holding the same level of strictness on privacy protection, DIP only entails an error of about 0.4% of the original sample mean, and the error is 13.08% and 4.69% for LRM and OPM, respectively. And the

amount of improvement of DIP over LRM and OPM is 3170.0% and 1072.5%. In other words, DIP achieves the highest accuracy while guaranteeing differential privacy.

For logistic regression in the second example, DIP entails a small error, suggesting that any conclusion based on DIP's private regression would remain nearly the same as the one based on the original data. In terms of statistical accuracy, DIP delivers a substantial improvement of 561.7% over LRM, whereas OPM fails to produce any meaningful results due to an inaccurately estimated probability of 1 or 0 across all settings.

For personalized recommendations in the third example, DIP performs the best and yields a significant improvement of 81.6%, 153.4%, and 7.8% over LRM, OPM, and EXM, respectively. This result implies that, while protecting data privacy, preserving the original data distribution also ensures high prediction accuracy. Additional supporting numerical evidence is given in Appendix S2.

3. Discussion

Differential privacy has become a standard of data privacy protection, as a large amount of sensitive information is collected and stored in a digital form. This paper proposes a novel privatization method, DIP, which preserves the original data's distribution while satisfying differential privacy. DIP mitigates the trade-off between data privacy and statistical accuracy. Consequently, any downstream privatized statistical analysis or machine learning task leads to the same conclusion as if the original data were used, which is a unique aspect that all existing mechanisms may not enjoy. Second, DIP is differentially private even if underlying data have unbounded support or unknown distributions. Our extensive numerical studies demonstrate the utility of DIP against top competitors across many distinct application scenarios. On the other hand, DIP requires a hold-out sample to estimate the data's distribution. This may reduce the released sample size due to sample splitting.

The proposed methodology also opens up several future fronts. One is the generalization to local differential privacy, which provides further privacy protection for data in a local device or server. Some discussions of DIP on local differential privacy are provided in Appendix S3. Another direction is the extension to independent but not identically distributed data, in which DIP can still ensure differential privacy while preserving the distribution, but requires repeated measurements for each individual. See Appendix S4 for more details.

4. Methods

4.1. Differential privacy

Differential privacy ensures that the substitution of any single observation in a dataset would have a small impact on the publicly shared information, which can be measured by ε-differential privacy. Suppose $\mathbf{Z}$ is a random sample from a cumulative distribution function (cdf) F, and $\mathtt{m}$ is a *privatization mechanism* which privatizes $\mathbf{Z}$ into $\tilde{\mathbf{Z}}$ for public release. Let $\mathbf{z}$ and $\mathbf{z}'$ be two *adjacent* realizations of $\mathbf{Z}$, which differ in only one observation. Then ε-differential privacy requires that the ratio of the probability of any privatized event given one sample to the probability of it given the other sample is upper bounded by e^{ε}, that is:

Definition 1 (*Dwork et al., 2006c; Dwork, 2006a*)**.** A privatization mechanism $\mathtt{m}(\cdot)$ satisfies ε-differential privacy if

$$\sup_{\mathbf{z},\mathbf{z}'} \sup_{B} \frac{P\big(\mathtt{m}(\mathbf{Z}) \in B | \mathbf{Z} = \mathbf{z}\big)}{P\big(\mathtt{m}(\mathbf{Z}) \in B | \mathbf{Z} = \mathbf{z}'\big)} \leq e^{\varepsilon},$$

where B is a measurable set and $\varepsilon \geq 0$ is a privacy factor that is usually small. For convenience, the ratio is defined as 1 when the numerator and denominator are 0.

Definition 1 requires that the ratio of conditional probabilities of any privatized event (i.e., the set B) given two *adjacent* data realizations is no greater than e^{ε}. Here ε is called a *privacy factor*, which characterizes the budget of privacy protection. For example, a small value of ε renders a strict privacy protection policy.

4.2. Implications of differential privacy

Differential privacy protects against any data identification or revealing data values of the original sample, by adversarial attack. However, the disclosure odds can increase as an adversarial attack repeats multiple times. Subsequently, to understand the level of protection of differential privacy, as specified by the privacy factor ε in Definition 1, we generalize Theorem 2.4 of Wasserman and Zhou (2010) to repeated adversarial attacks.

Consider that an adversary makes M queries to obtain M independent ε-differentially private samples, intending to reveal the values of an original sample $\mathbf{Z} = (Z_1, \ldots, Z_n)$. Lemma 1 details the level of protection by ε-differential privacy against data identification.

Lemma 1. *Suppose $\tilde{\mathbf{Z}}^{(1)}, \ldots, \tilde{\mathbf{Z}}^{(M)}$ are M independent yet ε-differentially private copies of $\mathbf{Z}$. Then any hypothesis test to identify the value of the i_0th observation Z_{i_0}, namely $H_0 : Z_{i_0} = \mu_0$ versus $H_1 : Z_{i_0} = \mu_1$, based on $\tilde{\mathbf{Z}}^{(1)}, \ldots, \tilde{\mathbf{Z}}^{(M)}$ has statistical power no greater than $\gamma e^{M\varepsilon}$ with any $\mu_1 \neq \mu_0$ and any $i_0 = 1, \ldots, n$, given a significance level $\gamma > 0$.*

In particular, Lemma 1 states that it is impossible to reject a null hypothesis H_0 that an observation equals a specific value μ_0 in any sample because of a small power $\gamma e^{M\varepsilon}$ for sufficiently small ε, especially so in a one-time data-release scenario with $M = 1$. However, the information leak could occur when M increases, while holding ε fixed in that H_0 is eventually rejected as a result of increased power.

In summary, ε-differential privacy only protects against data identification when M is limited. An adversarial attack will eventually break the defense of ε-differential privacy as the number of attacks M increases. In practice, M needs to be limited for ε-differential privacy depending on the privacy factor ε.

Next, we show, in Lemma 2, that any linear perturbation or transformation does not entail ε-differential privacy for the original data with unbounded support. This means that we must seek nonlinear transformations beyond the linear domain to protect data with unbounded support, which motivates the proposed method (1).

Lemma 2. *For any multivariate random sample $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, any linear privatization mechanism $\mathfrak{m}(\cdot)$: $\mathfrak{m}(\mathbf{Z}_i) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{Z}_i + \mathbf{e}_i$; $i = 1, \ldots, n$, is not ε-differentially private when $\mathbf{Z}_i$ has unbounded support, where $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ ($\boldsymbol{\beta}_1 \neq 0$) are any fixed coefficients, and $\mathbf{e}_i$ is any random noise vector.*

4.3. Theoretical justification

This subsection discusses DIP in the context of a known data distribution for motivation, which paves up the way for the theoretical justification of the distribution preservation property. Then, in a subsequent subsection, we further expand the method to an unknown data distribution by estimating it based on a hold-out sample.

Univariate continuous distributions. Suppose $(Z_1, \ldots, Z_n)$ is a random sample of a given cumulative distribution function F. We begin with our discussion with a known continuous F.

Our privatization proceeds by transforming each Z_i through the following steps, as displayed in Fig. 1; $i = 1, \ldots, n$. First, we apply F to Z_i to yield a uniformly distributed variable $F(Z_i)$. Second, we add an independent noise e_i to perturb $F(Z_i)$ for privacy protection, where e_i is randomly sampled from a Laplace distribution $Laplace(0, 1/\varepsilon)$. Here the scale parameter ε ensures that our privatization satisfies ε-differential privacy. Finally, we apply a nonlinear transformation H to produce a privatized sample that follows the original distribution F.

The specific form of H depends on the data type. For a univariate continuous sample, H consists of two parts. The cumulative distribution function G is applied to $F(Z_i) + e_i$ to produce a uniform variable, where G's explicit expression is given in Appendix S1.1. Then, the inverse cumulative distribution function F^{-1} is applied to obtain the original distribution. In summary, we generate a DIP's privatized sample $(\tilde{Z}_1, \ldots, \tilde{Z}_n)$ via a formula:

$$\tilde{Z}_i = H\big(F(Z_i) + e_i\big), \quad H(\cdot) = F^{-1} \circ G(\cdot), \tag{1}$$

where $\circ$ denotes function composition.

Univariate discrete distributions. For a discrete or mixed variable, that is, a variable with both continuous and discrete components, (1) still applies if we replace F by a smoothed F, a process we call continualization, as illustrated in Fig. 2. Specifically, we subtract a uniformly distributed random variable U_i from Z_i, such that the distribution of the modified Z_i follows the smoothed F, followed by privatization in (1). Then, we apply a ceiling function such that the privatized sample follows the original distribution F, c.f., the Appendices S1.3 and S1.4.

Multivariate distributions. Suppose $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ is a random sample of a multivariate distribution F, where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})$ is a p-dimensional vector. We privatize $\mathbf{Z}_i$ by applying (1), directly or after continualization, to each variable sequentially via the probability chain rule. In particular, we privatize its first component Z_{i1} to yield its privatized value $\tilde{Z}_{i1}$. Then privatize Z_{i2} given $\tilde{Z}_{i1}$ using the conditional cumulative distribution $F(Z_{i2}|Z_{i1})$ and the inverse function $F^{-1}(Z_{i2}|\tilde{Z}_{i1})$ to replace F and F^{-1} in (1), respectively, to yield $\tilde{Z}_{i2}$, and so forth. This leads to our general formula for DIP:

$$\tilde{Z}_{i1} = \mathfrak{m}_1(Z_{i1}), \quad \tilde{Z}_{il} = \mathfrak{m}_l(Z_{il}|\tilde{Z}_{i1}, \ldots, \tilde{Z}_{i,l-1}); \quad l = 2, \ldots, p, \tag{2}$$

where $\mathfrak{m}_l$ denotes (1) for the lth variable, with F in (1) (and the corresponding F^{-1}) replaced by the marginal distribution of Z_{i1} for $l = 1$, and the conditional distribution of Z_{il} given $\tilde{Z}_{i1}, \ldots, \tilde{Z}_{i,l-1}$ for $l = 2, \ldots, p$.

The properties for both univariate and multivariate privatizations are summarized in Theorem 1.

Theorem 1. *DIP in (2) is ε-differentially private, and the privatized sample follows the original distribution F when F is known.*

It is important to note that DIP in (1) and (2) differs substantially from a sampling method. A sampling method generates synthetic data from F, which does not require raw data $\mathbf{Z}$ when F is known. On the other hand, DIP preserves the data identifier i of $\mathbf{Z}_i$, $i = 1, \ldots, n$. This property is useful for data integration or personalization. For example, when collecting an additional column of data from the same individual i, for example, user i's ratings on a new movie in the context of Section 2.3.3, we may integrate them with $\tilde{\mathbf{Z}}_i$. In contrast, a sampling method does not retain any data identifier.

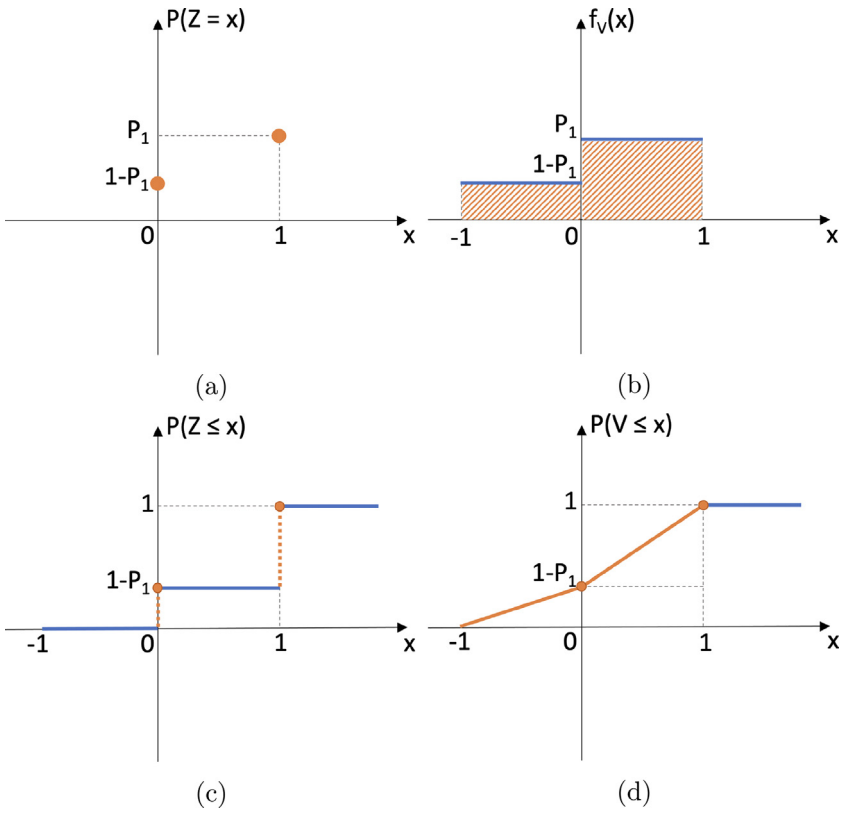

Fig. 2. An example of the univariate continualization method in DIP. (**a**) A variable Z follows a Bernoulli distribution, which takes the value 1 with probability P_1 and the value 0 with probability $1-P_1$. The two orange dots represent the two probability masses. (**b**) To continualize this distribution, we introduce $V = Z - U$, with U being an independent random variable following a uniform distribution between 0 and 1. The two probability masses $1-P_1$ and P_1 are now evenly spread on intervals $(-1, 0)$ and $(0, 1)$ (in the orange shade), respectively. And the probability density function of this new variable V is drawn in the blue lines. (**c**) The cumulative distribution function of Z has two jumps — one at 0, and the other at 1. And the heights of the jumps correspond to the probabilities of the two values. (**d**) After continualization, the cumulative distribution function of V becomes continuous, where the jumps are replaced by orange lines (i.e., linear interpolants).

4.4. The proposed method

In practice, F is usually unknown and has to be replaced by an estimate $\hat{F}$, for example, the empirical cumulative distribution function. However, $\hat{F}$ needs to be independent of the privatized data to satisfy differential privacy. Towards this end, we either (i) construct $\hat{F}$ based on a random subsample of $\mathbf{Z}$, called a *hold-out sample*. The hold-out sample is neither privatized nor released while the remaining sample is *to-be-privatized*, c.f., Appendix S1.2; or (ii) use a public dataset that follows the same distribution F as the hold-out sample, while treating the entire $\mathbf{Z}$ as the to-be-privatized sample. For example, the American Community Survey data are public and the U.S. census data are private, both coming from the same population. The former can serve as a hold-out sample for the latter. Here both the hold-out and the to-be-privatized samples are fixed once assigned. We then apply (2) with F replaced by $\hat{F}$ to the to-be-privatized sample, which asymptotically preserves the original distribution when $\hat{F}$ is a consistent estimate of F, for example, the smoothed empirical cumulative distribution function, c.f., Theorem 2. See Appendices S1.5 and S1.6 for the complete technical details of constructing the DIP mechanism for univariate and multivariate variables with $\hat{F}$, respectively.

Algorithm 1 summarizes the entire privatization process of DIP. Let $N = n + m$ be the total sample size, where m is the hold-out sample size. Proposition 1 concerns the computational efficiency and scalability of Algorithm 1.

Proposition 1. *The computational complexity of Algorithm 1 is* $O(pN \log N)$.

Theorem 2 establishes DIP's ε-differential privacy and its asymptotic distribution preservation as the size of the hold-out sample tends to infinity.

Theorem 2. *DIP in Algorithm 1 is ε-differentially private, and the privatized sample follows the original distribution F asymptotically as $m \to \infty$ when F is unknown.*

Algorithm 1 Distribution-invariant privatization

Input: A to-be-privatized sample $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, a hold-out sample, dimension p, the privacy factor ε.
for $i = 1, \ldots, n$ **do**
  for $l = 1, \ldots, p$ **do**
    Continualize Z_{il} if it is not continuous.
    Privatize Z_{il} into $\tilde{Z}_{il}$ following (2) with a privacy factor ε/p and with $\hat{F}$ estimated by the hold-out sample.
  end for
  for $l = 1, \ldots, p$ **do**
    Apply a corresponding ceiling function to $\tilde{Z}_{il}$ if Z_{il} is not continuous.
  end for
end for
Output: A privatized sample $(\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n)$.

To guarantee ε-differential privacy, we apply the sequential composition (Dwork et al., 2006c; Kairouz et al., 2017) and require each $\mathfrak{m}_l$ in (2) to be ε/p-differentially private, which is enforced by sampling e_{il} from a Laplace distribution $Laplace(0, p/\varepsilon)$. A small loss of statistical accuracy may incur, depending on the estimation precision of $\hat{F}$ for F in a finite-sample situation. While a large m provides a refined estimate of F, a large n renders a large to-be-privatized sample for downstream statistical analysis. In general, one may choose a reasonable m to retain statistical accuracy. However, when $N = n + m$ is fixed, one needs to consider if the released sample size is adequate.

As a technical note, the DIP result continues to hold for any consistent estimator of F, which is useful when additional information is available. For example, if Z_i's follow a normal distribution $N(\mu, 1)$ with an unknown mean value μ, then $\hat{F}$ can be chosen as the continuous cdf of a normal distribution with an estimated mean $\hat{\mu}$ and variance 1.

It is worth noting that the privacy of raw data in the hold-out sample is also protected when a public dataset is unavailable for estimating F. First, any transmission, alteration, querying, or release of any raw data in the hold-out sample is not permissible. Second, only a continuous version of the estimated F is constructed, which guarantees that $\hat{F}$ does not contain any probability mass of the raw data. This is achieved through adding noise to the raw data in the hold-out sample. See Appendices S1.5 and S1.6 for more details. Therefore, the privacy protection of a hold-out sample remains a high standard.[1] As shown in Lemma 3, the probability of identifying a continualized value of a hold-out sample from the privatized data is zero. In other words, there is no privacy leakage of the hold-out sample.

Lemma 3. *For any individual j in the hold-out sample, that is, $j = n + 1, \ldots, n + m$, let $\mathbf{V}_j$ be the continualized version of $\mathbf{Z}_j$. Then $P(\mathbf{V}_j = \boldsymbol{v} | \tilde{\mathbf{Z}}_1 = \tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{Z}}_n = \tilde{\mathbf{z}}_n) = 0$ for any $\boldsymbol{v}, \tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n \in \mathbb{R}^p$.*

The notion of "hold-out" is an analogy to that of retaining sensitive data confidential, as required by the National Institutes of Health (NIH) to perform medical research. In particular, NIH's policy states that personally identifiable, sensitive, and confidential information should "not be housed on portable electronic devices", and researchers should "limit access to personally identifiable information through proper access controls."[2] Accordingly, the privacy of the hold-out sample here is protected in the same fashion — ideally stored in a room without Internet and prevented from unauthorized access, in addition to noise injection, continualization, and the requirement of not being transmitted, altered, queried, or released.

Acknowledgments

The authors thank the editor and three reviewers for insightful comments and suggestions, which improve the article significantly. This research is supported in part by NSF, USA grant DMS-1952539, and NIH, USA grants R01AG069895, R01AG065636, 1R01GM126002, R01HL105397, R01AG074858, and U01AG073079.

Appendix A. Supplementary materials

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2022.05.004. The Appendix contains the technical details of the proposed method, details about experimental setup and additional numerical studies, the connection of the proposed method with local differential privacy, generalization of the proposed method to independent but not identically distributed data, technical proofs, and additional figures and tables.

[1] Differential privacy as in Definition 1 is not well defined on the hold-out sample due to no adjacent realizations being permissible.

[2] The original policy is available at https://grants.nih.gov/grants/policy/nihgps/html5/section_2/2.3.12_protecting_sensitive_data_and_information_used_in_research.htm.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L., 2016. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 308–318.

Agarwal, N., Suresh, A.T., Yu, F.X., Kumar, S., McMahan, H.B., 2018. cpSGD: Communication-efficient and differentially-private distributed SGD. In: Advances in Neural Information Processing Systems. pp. 7564–7575.

Apple Differential Privacy Team, 2017. Learning with privacy at scale. Apple Mach. Learn. J. 1 (8).

Avella-Medina, M., 2021. Privacy-preserving parametric inference: A case for robust statistics. J. Amer. Statist. Assoc. 116 (534), 969–983.

Bowen, C.M., Liu, F., 2020. Comparative study of differentially private data synthesis methods. Statist. Sci. 35 (2), 280–307.

Cai, T.T., Wang, Y., Zhang, L., 2019. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. arXiv preprint arXiv:1902.04495.

Chaudhuri, K., Monteleoni, C., Sarwate, A.D., 2011. Differentially private empirical risk minimization. J. Mach. Learn. Res. 12 (Mar), 1069–1109.

Chen, Y., Machanavajjhala, A., Reiter, J.P., Barrientos, A.F., 2016. Differentially private regression diagnostics. In: IEEE 16th International Conference on Data Mining. pp. 81–90.

Cohen, A., Nissim, K., 2020. Towards formalizing the GDPR's notion of singling out. Proc. Natl. Acad. Sci. 117 (15), 8344–8352.

Day One Staff, 2018. Protecting data privacy: How Amazon is advancing privacy-aware data processing. Available at https://blog.aboutamazon.com/amazon-ai/protecting-data-privacy.

Ding, B., Kulkarni, J., Yekhanin, S., 2017. Collecting telemetry data privately. In: Advances in Neural Information Processing Systems. pp. 3571–3580.

Dong, J., Roth, A., Su, W.J., 2019. Gaussian differential privacy. arXiv preprint arXiv:1905.02383.

Duchi, J.C., Jordan, M.I., Wainwright, M.J., 2018. Minimax optimal procedures for locally private estimation. J. Amer. Statist. Assoc. 113 (521), 182–201.

Durfee, D., Rogers, R.M., 2019. Practical differentially private top-k selection with pay-what-you-get composition. In: Advances in Neural Information Processing Systems. pp. 3527–3537.

Dwork, C., 2006a. Differential privacy. In: The 33rd International Colloquium on Automata, Languages and Programming. Springer, pp. 1–12.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M., 2006b. Our data, ourselves: Privacy via distributed noise generation. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques. pp. 486–503.

Dwork, C., McSherry, F., Nissim, K., Smith, A., 2006c. Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd Theory of Cryptography Conference. pp. 265–284.

Dwork, C., Roth, A., 2014. The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. 9 (3–4), 211–407.

Dwork, C., Rothblum, G.N., Vadhan, S., 2010. Boosting and differential privacy. In: 2010 IEEE 51st Annual Symposium on Foundations of Computer Science. pp. 51–60.

Erlingsson, U., Pihur, V., Korolova, A., 2014. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. ACM Press, http://dx.doi.org/10.1145/2660267.2660348.

Evfimievski, A., Gehrke, J., Srikant, R., 2003. Limiting privacy breaches in privacy preserving data mining. In: Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. pp. 211–222.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9 (3), 432–441.

Funk, S., 2006. Netflix update: Try this at home. URL http://sifter.org/~simon/journal/20061211.html.

Gong, R., Meng, X.-L., 2020. Congenial differential privacy under mandated disclosure. In: Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference. pp. 59–70.

Goroff, D.L., 2015. Balancing privacy versus accuracy in research protocols. Science 347 (6221), 479–480.

Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C.S., et al., 2020. Transparency and reproducibility in artificial intelligence. Nature 586 (7829), E14–E16.

Hall, R., Rinaldo, A., Wasserman, L., 2012. Random differential privacy. J. Priv. Confid. 4 (2), 43–59.

Hall, R., Rinaldo, A., Wasserman, L., 2013. Differential privacy for functions and functional data. J. Mach. Learn. Res. 14 (Feb), 703–727.

Han, T., Nebelung, S., Haarburger, C., Horst, N., Reinartz, S., Merhof, D., Kiessling, F., Schulz, V., Truhn, D., 2020. Breaking medical data sharing boundaries by using synthesized radiographs. Sci. Adv. 6 (49), eabb7973.

Harper, F.M., Konstan, J.A., 2015. The MovieLens datasets: History and context. ACM Trans. Interact. Intell. Syst. 5 (4), 1–19.

Hie, B., Cho, H., Berger, B., 2018. Realizing private and practical pharmacological collaboration. Science 362 (6412), 347–350.

Jobin, A., Man, K., Damasio, A., Kaissis, G., Braren, R., Stoyanovich, J., Van Bavel, J.J., West, T.V., Mittelstadt, B., Eshraghian, J., et al., 2021. AI reflections in 2020. Nat. Mach. Intell. 3 (1), 2–8.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science 349 (6245), 255–260.

Kairouz, P., Oh, S., Viswanath, P., 2017. The composition theorem for differential privacy. IEEE Trans. Inform. Theory 63 (6), 4037–4049.

Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F., 2020. Secure, privacy-preserving and federated machine learning in medical imaging. Nat. Mach. Intell. 2 (6), 305–311.

Karwa, V., Slavković, A., 2016. Inference using noisy degrees: Differentially private β-model and synthetic graphs. Ann. Statist. 44 (1), 87–112.

Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A., 2011. What can we learn privately? SIAM J. Comput. 40 (3), 793–826.

Kearns, M., Roth, A., Wu, Z.S., Yaroslavtsev, G., 2016. Private algorithms for the protected in social network search. Proc. Natl. Acad. Sci. 113 (4), 913–918.

Kenthapadi, K., Tran, T.T.L., 2018. Pripearl: A framework for privacy-preserving analytics and reporting at LinkedIn. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 2183–2191.

Lin, S.P., Perlman, M.D., 1985. A Monte Carlo comparison of four estimators of a covariance matrix. Multivariate Anal. 6, 411–429.

McSherry, F., Talwar, K., 2007. Mechanism design via differential privacy. In: 48th Annual IEEE Symposium on Foundations of Computer Science. pp. 94–103.

Mironov, I., 2017. Rényi differential privacy. In: 2017 IEEE 30th Computer Security Foundations Symposium. pp. 263–275.

Moro, S., Cortez, P., Rita, P., 2014. A data-driven approach to predict the success of bank telemarketing. Decis. Support Syst. 62, 22–31.

Narayanan, A., Shmatikov, V., 2008. Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy, Sp 2008. IEEE, pp. 111–125.

Nayak, C., 2020. New privacy-protected Facebook data for independent research on social media's impact on democracy. Available at https://research.fb.com/blog/2020/02/new-privacy-protected-facebook-data-for-independent-research-on-social-medias-impact-on-democracy.

Rohde, A., Steinberger, L., 2020. Geometrizing rates of convergence under local differential privacy constraints. Ann. Statist. 48 (5), 2646–2670.

Santos-Lozada, A.R., Howard, J.T., Verdery, A.M., 2020. How differential privacy will affect our understanding of health disparities in the United States. Proc. Natl. Acad. Sci. 117 (24), 13405–13412.

Schum, D.A., 2001. The Evidential Foundations of Probabilistic Reasoning. Northwestern University Press.

United States Census Bureau, 2020. Disclosure avoidance and the 2020 census. Available at https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html.

University of California, 2014. Annual report on employee compensation. Available at http://compensation.universityofcalifornia.edu/payroll2010.

Vadhan, S., 2017. The complexity of differential privacy. In: Tutorials on the Foundations of Cryptography. Springer, pp. 347–450.

Venkatramanan, S., Sadilek, A., Fadikar, A., Barrett, C.L., Biggerstaff, M., Chen, J., Dotiwalla, X., Eastham, P., Gipson, B., Higdon, D., et al., 2021. Forecasting influenza activity using machine-learned mobility map. Nature Commun. 12 (1), 1–12.

Wasserman, L., Zhou, S., 2010. A statistical framework for differential privacy. J. Amer. Statist. Assoc. 105 (489), 375–389.