Journal of Big Data

RESEARCH Open Access

Have the cake and eat it too: Differential Privacy enables privacy and precise analytics

Rishabh Subramanian[1*]

*Correspondence:
rishabhs@uchicago.edu

[1] University of Chicago, Chicago, USA

Abstract

Existing research in differential privacy, whose applications have exploded across functional areas in the last few years, describes an intrinsic trade-off between the privacy of a dataset and its utility for analytics. Resolving this trade-off critically impacts potential applications of differential privacy to protect privacy in datasets even while enabling analytics using them. In contrast to the existing literature, this paper shows how differential privacy can be employed to *precisely—not approximately—retrieve* the analytics on the original dataset. We examine, conceptually and empirically, the impact of noise addition on the quality of data analytics. We show that the accuracy of analytics following noise addition increases with the privacy budget and the variance of the independent variable. Also, the accuracy of analytics following noise addition increases disproportionately with an increase in the privacy budget when the variance of the independent variable is greater. Using actual data to which we add Laplace noise, we provide evidence supporting these two predictions. We then demonstrate our *central thesis* that, once the privacy budget employed for differential privacy is declared and certain conditions for noise addition are satisfied, the slope parameters in the original dataset can be *accurately retrieved* using the estimates in the modified dataset of the variance of the independent variable and the slope parameter. Thus, differential privacy can enable robust privacy as well as *precise* data analytics.

Keywords: Data analytics, Data mining, Data privacy, Differential privacy, DiD, Difference-in-difference, OLS, Ordinary least squares, Prediction, Regression

Introduction

> *"Information is the oil of the 21st century, and analytics is the combustion engine."—Peter Sondergaard, Senior Vice President and Global Head of Research at Gartner, Inc.*

As the above quote highlights, data analysis and prediction have become the cornerstone of corporate and public policy. While powerful insights can be obtained when granular data—often about individuals—are shared for research, concerns about the privacy of such granular data limit society's potential to put it to optimal use. Individuals' privacy can get compromised even when their data is shared with the individual's approval and is stripped of personal identifiers. Na et al. [1] Show how researchers could re-identify

95 percent of individual adults from the National Health and Nutrition Examination Survey using machine learning techniques; [2] similarly show a high re-identification rate of data. A prominent example is of Governor William Weld, the former Governor of Massachusetts, who was subjected to such re-identification using a linkage attack [3].

Differential privacy has emerged as a technique to ensure privacy of individuals in a dataset, even when their data is shared publicly [4]. As differential privacy changes the *process* for accessing data, rather than the database itself, it enables individuals' privacy even when the data is subjected to various attacks on privacy. The use of differential privacy by the 2020 U.S. Census signals a seminal change in government statistics [5]. Leading corporations and governments have started employing differential privacy into their datasets; see [5–7].

In the last few years, there has been an explosion of research articles that apply differential privacy to various functional areas such as healthcare [8–22], learning [23–38], location-based services [39–47], internet-based collaboration [48], Internet of Things [49–51], block-chains [52–54], cyber-physical systems [55–58], neural networks [59], social media and social network analysis [60–62], crowd-sourcing [63–65], and mobile edge computing environments [66, 67]. Pejó and Desfontaines [68] study the numerous variants and extensions to adapt differential privacy to different scenarios and attacker models.

Existing research in this area emphasizes an intrinsic trade-off between the privacy of a dataset and its utility for analytics. In their survey of the privacy literature, [69] describe this trade-off as "differential privacy provides either very little privacy or very little utility or neither." In contrast to such existing literature, this paper shows that differential privacy can be employed to *precisely—not approximately—retrieve* the associations in the original dataset. As viable methods for protection of privacy that do not impinge on the quality of data analytics are cardinal to our increasingly data-reliant and privacy-conscious society, our study makes an *important contribution* by highlighting that differential privacy can enable privacy while simultaneously preserving the quality of data analytics as in the original data.

We examine, conceptually and empirically, the impact of noise addition using differential privacy on the quality of data analytics on a modified dataset, i.e. a dataset with noise. As associations between the dependent and independent variables are typically captured using the slope parameter in a regression, we examine the impact of noise addition on the slope parameter. We obtain two key results. First, the accuracy of analytics following noise addition increases with the privacy budget and the variance of the independent variable. Second, the accuracy of analytics following noise addition increases disproportionately with an increase in the privacy budget when the variance of the independent variable is greater. To test these two predictions, we use actual data, where both the dependent and explanatory variables are private. We add Laplace noise to both these variables and then compare the slopes in the original and modified datasets to provide evidence supporting these two predictions. We thus conceptually and empirically establish that the utility-privacy trade-off exists in differential privacy.

We then ask the *central question* in this study: Can this utility-privacy trade-off be overcome using differential privacy? We highlight that differential privacy can ensure

precise data analytics even while preserving the privacy of the individuals in a dataset, provided the noise added satisfies the following criteria. If the dependent variable and an explanatory variable are both private variables, three conditions must be satisfied. First, noise added to dependent variable is independent of the explanatory variable. Second, noise added to the explanatory variable is independent of the dependent variable. Third, these noises are, in turn, independent of each other. Given these criteria, we show that once the privacy budget employed to construct a differentially private dataset is declared, the original slope parameter can be *precisely retrieved* using the variance of the independent variable and the slope parameter estimated using the modified dataset. Critically, we demonstrate these results by being agnostic about the nature of the statistical distribution from which the noise is added to achieve differential privacy. As revealing the privacy budget used to arrive at the differentially private dataset does not necessarily compromise the privacy of the dataset, differential privacy can enable us to overcome the utility-privacy trade-off.

If only the dependent variable is private while the explanatory variable is a public variable, noise needs to be added to only the dependent variable. In this case, if noise added to dependent variable is independent of the explanatory variable, the original slope parameter is identical to the estimate generated using the modified dataset; this result is again agnostic to the nature of the statistical distribution from which the noise is added to achieve differential privacy.

Our study makes an *important contribution* to the differential privacy literature. In their survey of the privacy literature, [69] classify differential privacy, k-anonymity, l-diversity and t-closeness as the techniques that employ input privacy for data mining. Outlining the advantage of differential privacy through the contributions of [10, 70, 69] highlight that "differential privacy is becoming a popular research area as it guarantees data privacy... (and) ensures utility as noise addition is minimal thus *providing a close approximation of original results*." (emphasis added) However, outlining its disadvantages, they write "differential privacy provides either very little privacy or very little utility or neither." A similar belief was expressed in [71], where they mention "It is believed that certain paradigms such as differential privacy reduce the information content too much to be useful in practical situations." (pp. 322) In contrast, our study shows that by declaring the privacy budget used in generating a differentially private dataset, *precise—not approximate as claimed in* [69]*—data analytics can be performed using the modified dataset even while preserving its privacy*.

Within the scope of the utility-privacy trade-off, our study contrasts:

1. The claim in [69] that "differential privacy provides either very little privacy or very little utility or neither." Our study shows that both privacy and utility can be obtained using differential privacy.
2. The thesis in [72] that techniques for privacy preservation have "a noticeable impact of privacy-preservation techniques in predictive performance." Our study shows that differential privacy can ensure *no noticeable impact* of privacy-preservation techniques in predictive performance.

3. The concern raised in [5] with respect to the use of differential privacy by the 2020 U.S. Census that "transition to differential privacy has raised a number of questions about the proper balance between privacy and accuracy in official statistics." Our study shows that these concerns about the balance between privacy and accuracy—with respect to analytics using the census data—may be misplaced.

Our study also contributes to the literature on privacy preserving data analytics. Zhang et al. [73] survey the literature on privacy preserving association rule mining, especially focusing on the present methodologies for the same. Ahluwalia et al. [74] study association rule mining where mining is conducted by a third party over data located at a central location is updated from several source locations. We show that differential privacy can be used to completely preserve the utility of data analytics, while ensuring the privacy of data.

The paper is structured as follows. Section analyzes the effects of noise addition on the accuracy of analytics. Section postulates the key result in our study. Section concludes the paper.

Effect of noise addition using differential privacy on data analytics

Following [4, 75], ϵ-differential privacy is defined formally as follows. If ϵ is a positive real constant, A is a randomised process, D and D$'$ are databases that differ by the data of one individual, and O is some output of the process A, then ϵ-differential privacy is defined as:

$$[A(D) = O] \leq e^{\epsilon} \cdot P[A(D') = O] \tag{1}$$

The smaller ϵ is, the closer the probabilities above are, and, therefore, the more differentially private the process is. Conversely, a higher ϵ implies a less differentially private process.

Having defined ϵ-differential privacy, we now study the central thesis of this paper: the purported trade-off between utility and privacy of a differentially private dataset. As the correlations between dependent and independent variables—in univariate or multivariate settings—are most important in data analytics, we study the effect of adding noise to enable differential privacy on the correlations as measured by the slope parameter in a regression.

We first consider the case where both the dependent variable y and the independent variable x are private.

Adding noise to private dependent and independent variables: conceptual analysis

Denote $\Lambda(\mu, \sigma)$ as a function that finds a random value from a distribution with mean μ and standard deviation σ. We use $\Lambda'(\mu, \sigma)$ to denote a random draw that is different from $\Lambda(\mu, \sigma)$. We add noise from a distribution with $\sigma = \frac{\alpha}{\epsilon}$, where α is a constant and ϵ is the privacy parameter, to both dependent and independent variables in an ordinary least squares (OLS) regression.[1] We get the following equation:

[1] α equals $\sqrt{2}$ for a Laplace distribution for instance.

$$\left[y_i + \Lambda'\left(\mu, \frac{\alpha}{\epsilon}\right)\right] = \beta_0 + \beta_1 \cdot \left[x_i + \Lambda\left(\mu, \frac{\alpha}{\epsilon}\right)\right] + v_i, \tag{2}$$

where the estimate of β_1 is given by:

$$\beta_1 = \frac{covar(x, y)}{var(x)}, \tag{3}$$

where $covar(x, y)$ denotes the covariance between random variables x and y and $var(x)$ denotes the variance of random variable x. Similarly, after noise addition, the new slope parameter β_1' equals:

$$\beta_1' = \frac{covar\left\{x + \Lambda\left(\mu, \frac{\alpha}{\epsilon}\right), y + \Lambda'\left(\mu, \frac{\alpha}{\epsilon}\right)\right\}}{var\left\{x + \Lambda\left(\mu, \frac{\alpha}{\epsilon}\right)\right\}} \tag{4}$$

The denominator equals

$$var\left\{x + \Lambda\left(\mu, \frac{\alpha}{\epsilon}\right)\right\} = var(x) + var\left\{\Lambda\left(\mu, \frac{\alpha}{\epsilon}\right)\right\} + 2 \cdot covar\left\{x, \Lambda\left(\mu, \frac{\alpha}{\epsilon}\right)\right\} \tag{5}$$

Now, as $\Lambda(\mu, \frac{\alpha}{\epsilon})$ is independent of x, the covariance between them equals 0 while variance of $\Lambda(\mu, \frac{\alpha}{\epsilon})$ equals $\frac{\alpha^2}{\epsilon^2}$. Therefore, the denominator simplifies as:

$$var\left\{x + \Lambda\left(\mu, \frac{\alpha}{\epsilon}\right)\right\} = var(x) + \frac{\alpha^2}{\epsilon^2} \tag{6}$$

The numerator equals

$$\begin{aligned} covar\left\{x + \Lambda\left(\mu, \frac{\alpha}{\epsilon}\right), y + \Lambda'\left(\mu, \frac{\alpha}{\epsilon}\right)\right\} &= covar(x, y) + covar\left\{x, \Lambda'\left(\mu, \frac{\alpha}{\epsilon}\right)\right\} \\ &+ covar\left\{\Lambda\left(\mu, \frac{\alpha}{\epsilon}\right), y\right\} + covar\left\{\Lambda\left(\mu, \frac{\alpha}{\epsilon}\right), \Lambda'\left(\mu, \frac{\alpha}{\epsilon}\right)\right\} \end{aligned} \tag{7}$$

We use the fact that $\Lambda(\mu, \frac{\alpha}{\epsilon})$ and $\Lambda'(\mu, \frac{\alpha}{\epsilon})$ are independent of each other, y is independent of $\Lambda(\mu, \frac{\alpha}{\epsilon})$, and x is independent of $\Lambda'(\mu, \frac{\alpha}{\epsilon})$. So, except the first covariance, the other three covariances are zero. So, the numerator simplifies as:

$$covar\left\{x + \Lambda\left(\mu, \frac{\alpha}{\epsilon}\right), y + \Lambda'\left(\mu, \frac{\alpha}{\epsilon}\right)\right\} = covar(x, y)$$

Using the simplified numerator and denominator, we get

$$\beta_1' = \frac{cov(x, y)}{var(x) + \frac{\alpha^2}{\epsilon^2}} \tag{8}$$

Using Eqs. (3) and (8) and denoting $var(x)$ as σ_x^2:

$$\frac{\beta_1}{\beta_1'} = 1 + \frac{\alpha^2}{\epsilon^2 \sigma_x^2} \tag{9}$$

Clearly, $\beta_1' < \beta_1$, which leads to the following result:

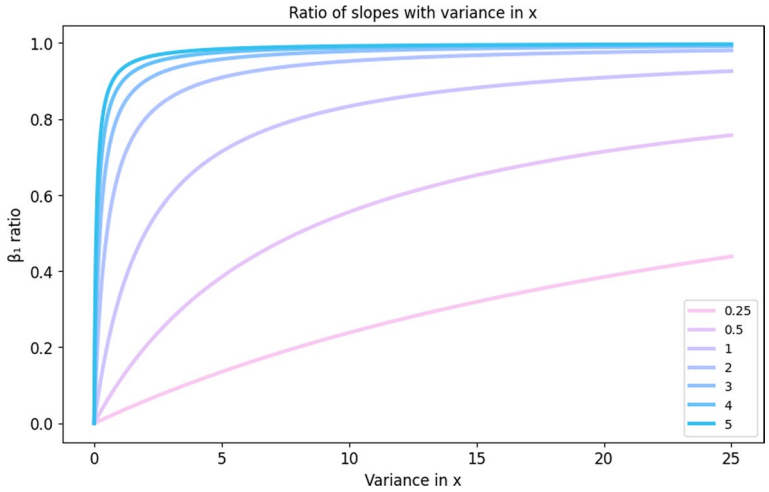

Fig. 1 Change in accuracy of analytics following noise addition

Result 1 If the dependent variable y and an explanatory variable x are both private variables, then the slope parameter used in data analytics is lower in magnitude after noise addition when compared to the slope parameter in the original dataset.

Figure 1 shows how the ratio $\frac{\beta_1'}{\beta_1}$ varies with the variance of x over different values for ϵ.

In Fig. 1 and in (9), we notice that, for a given value of ϵ, as the variance in x increases, β_1' approaches β_1. Similarly, for a given value of σ_x^2, β_1' approaches β_1 as ϵ increases. Thus, we get the following two results:

Result 2A If the dependent variable y and an explanatory variable x are both private variables, then the accuracy of analytics following noise addition increases with increases in the privacy budget (ϵ) and in the variance of the independent variable (σ_x^2).

$$\frac{d(\beta_1'/\beta_1)}{d\epsilon} > 0, \ \frac{d(\beta_1'/\beta_1)}{d(\sigma_x^2)} > 0. \tag{10}$$

Result 3A If the dependent variable y and an explanatory variable x are both private variables, then the accuracy of analytics following noise addition increases disproportionately with increase in privacy budget (ϵ) when the variance of the independent variable (σ_x^2) increases:

$$\frac{d^2(\beta_1'/\beta_1)}{d\epsilon \cdot d(\sigma_x^2)} > 0. \tag{11}$$

Adding noise to private dependent and independent variables: empirical evidence

In this section, we analyze the effect of noise addition to satisfy differential privacy to private dependent and independent variables on the accuracy of data analytics. We focus on two popular techniques used for analysis of data: ordinary least squares (OLS) regression and difference-in-difference estimation using panel data techniques ([76], Ch. 5, [77]). The estimate of the slope parameter in a difference-in-difference analysis

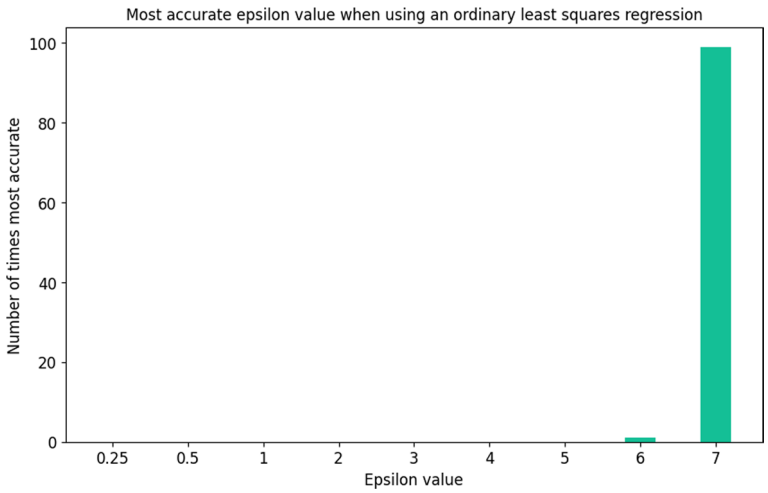

Fig. 2 Most accurate ϵ for OLS regression

resembles the discrete equivalent of a second-order derivative of the dependent variable w.r.t. the independent variable. Therefore, the variance of the independent variable in a difference-in-difference analysis is significantly greater than the variance of the independent variable in an OLS analysis. Thus, the simultaneous use of OLS and difference-in-difference estimation on the same dataset enables us to proxy the changes in the variance of the independent variable. In contrast, using two different datasets to proxy the changes in the variance of the independent variable would introduce other differences that may confound the empirical analysis. Thus, in our empirical analysis, the OLS estimates proxy the dataset with a lower variance for the independent variable while the difference-in-difference estimates proxy the dataset with a higher variance for the same.

We add noise from a Laplace distribution to a real dataset containing data on vaccination and health outcomes across all the states in the United States. The data on health outcomes—which includes the total number of cases per million people, the total number of deaths per million people, and the percentage case fatality rate—is collected from Oxford University's COVID-19 Government Response Tracker. The data on vaccination is collected from ourworldindata.org. The time period of the data is from Jan-2021, when vaccination first began in the U.S. to Apr-2021, when we had collected the data. We add noise to this data using nine different values of epsilon: 0.25, 0.5, 1, 2, 3, 4, 5, 6, and 7.

In both analyses, OLS and difference-in-difference, we compare the slope parameter we obtain on the noisy dataset (β_1') with the results obtained on the original dataset (β_1). We do this comparison for each value of epsilon to find the most accurate epsilon, the one with the least difference in the slope parameters vis-à-vis the original. We repeat this 100 times and aggregate the results to find the most accurate epsilon value. For each epsilon value, we also find the average of the squared difference in the slope parameters (β_1' and β_1) over the 100 repetitions to find the effects of noise addition.

The results of our analysis are shown in Figs. 2 and 3, which display a bar chart with the number of times each epsilon value was most accurate for OLS and difference-in-difference respectively. These figures confirm our theoretical prediction in Result 2A that

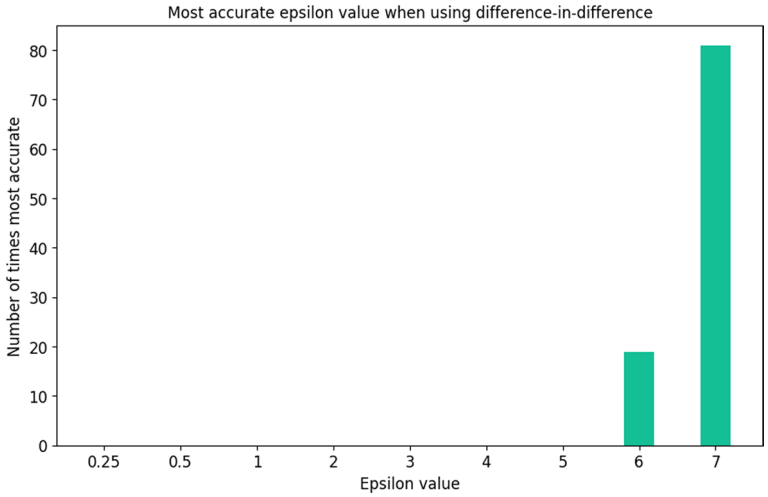

Fig. 3 Most accurate ϵ for difference-in-difference estimates

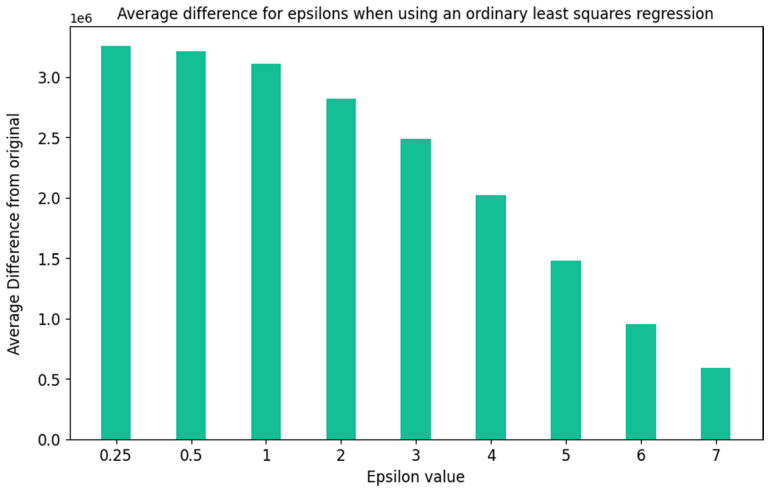

Fig. 4 Accuracy of analytics using ols regression for varying ϵ

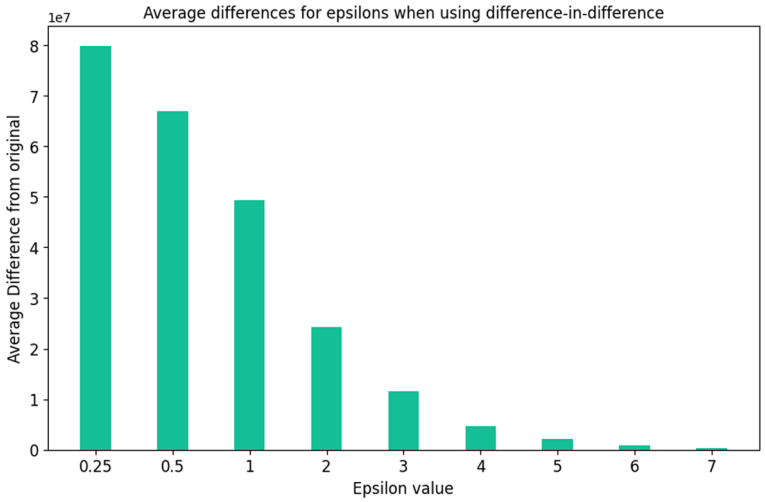

Fig. 5 Accuracy of Analytics using Difference-in-difference Estimates for Varying ϵ

Table 1 Average square of differences in slope parameter in original dataset and in dataset with noise addition using ordinary least squares (OLS) and difference-in-difference (DiD)

ϵ	OLS	DiD	ε	OLS	DiD
0.25	3.3	79.9	4	2.0	4.7
0.5	3.2	66.9	5	1.5	2.2
1	3.1	49.3	6	0.9	0.9
2	2.8	24.3	7	0.6	0.4
3	2.5	11.6			

an increase in the value of ϵ increases the accuracy of data analytics. Thus, we state the following result from our empirical analysis:

Result 2B As theoretically predicted in Result 2A, the empirical analysis using actual data confirms that the accuracy of analytics following noise addition increases with increase in the value of ϵ.

Figures 4 and 5 display a bar chart with the average sum of squares error in the slope parameters for the OLS and difference-in-difference respectively for each value of epsilon. We observe that the average difference over 100 iterations between the result obtained using the noisy dataset and the original dataset monotonically decreases with an increase in epsilon. This result is consistent with what we found using the earlier Figs. 2 and 3 charts that displayed the most accurate epsilon.

Table 1 displays the exact values for the average squares of differences between the slope parameters. We observe that as the value of epsilon rises, the average squares of differences falls monotonically—by approximately 6 times for ordinary least squares and approximately 200 times for difference-in-difference. Thus, there is a disproportionately larger drop in the average differences for the difference-in-difference estimate when compared to the estimates obtained using the ordinary least squares. is consistent with that of the theoretical analysis. Result 3A predicts that as the variance of the independent variable increases, an increase in ϵ disproportionately increases the accuracy. Therefore, we find that the empirical evidence when comparing the change in accuracy with ϵ for the difference-in-difference analysis versus that for the ordinary least squares analysis is consistent with Result 3A:

Result 3B As theoretically predicted in Result 3A, the effect of an increase in the value of ϵ is disproportionately more in a difference-in-difference analysis than in an ordinary least squares regression.

Thus, the empirical analysis clearly confirms that as the privacy budget (ϵ) increases, the utility of the data analytics—as measured by the slope parameter capturing the association between the dependent and independent variables—declines. On the other hand, the definition of differential privacy as in equation (1) shows clearly that as the privacy budget (ϵ) decreases, the data becomes more private. Thus, our conceptual and empirical analysis clearly demonstrates the intrinsic trade-off between privacy and utility when employing differential privacy. This trade-off has been described in [69], who outline

the disadvantages of differential privacy when they note that "differential privacy provides either very little privacy or very little utility or neither." Similarly, [72] highlight that techniques for privacy preservation have "a noticeable impact of privacy-preservation techniques in predictive performance."

Adding noise to only private dependent variable: conceptual analysis

Having analysed the case where both the dependent and independent variables are private, we now examine the impact on data analytics in the case where the independent variable is a public variable and so there is no need to add noise to the same to preserve privacy. As the case where the dependent variable is a public variable is not interesting from the perspective of data analysis, we ignore that case; we note, however, from the formula for the slope parameter in Eq. (3) that when no noise is added to the dependent variable, the slope parameter remains unchanged. In this case, we get the following equation:

$$\left[y_i + \Lambda'\left(0, \frac{\alpha}{\epsilon}\right)\right] = \beta_0 + \beta_1 \cdot x_i + v_i, \tag{12}$$

After noise addition, the new slope parameter β_1' equals:

$$\beta_1' = \frac{cov\left(x, y + \Lambda'\left(0, \frac{\alpha}{\epsilon}\right)\right)}{var(x)} \tag{13}$$

Replicating the steps as shown in section , we find that

$$cov\left\{x, y + \Lambda'\left(0, \frac{\alpha}{\epsilon}\right)\right\} = cov(x, y)$$

Therefore, in the case where the independent variable is a public variable,

$$\beta_1' = \beta_1 \tag{14}$$

This leads to our next result:

Result 4 When the independent variable is a public variable, the slope parameter remains unchanged after noise addition.

Key advantage of differential privacy for data analytics: precise analytics without losing privacy

Having conceptually and empirically demonstrated the trade-off between utility of differential privacy for data analytics and its ability to preserve privacy, we ask the *central question in this study*: Can this trade-off be avoided? As the slope parameter remains unchanged in the case where the independent variable is a public variable, we focus only on the case where both the dependent and independent variables are private. In this case, we highlight that differential privacy ensures the *precision* of data analytics even while preserving the privacy of the individuals in a dataset.

Combining Eqs. (6) and (9), the slope parameter in the original dataset can be regenerated from the slope parameter in the modified dataset using the variance of the independent variable in the modified dataset $\sigma_{x'}^2$ and the privacy budget ϵ as follows:

$$\beta_1 = \beta_1' \left\{ \frac{\epsilon^2 \sigma_{x'}^2}{\epsilon^2 \sigma_{x'}^2 - \alpha^2} \right\} \tag{15}$$

Thus, given the level of differential privacy employed in the modified dataset, i.e. with noise addition, the original slope parameter can be accurately retrieved using the variance calculated for the independent variable in the modified dataset $\sigma_{x'}^2$ and the slope parameter estimated in the modified dataset β_1' provided the following criteria are satisfied. If the dependent variable and an explanatory variable are both private variables, three conditions must be satisfied. First, noise added to dependent variable is independent of the explanatory variable. Second, noise added to the explanatory variable is independent of the dependent variable. Third, these noises are, in turn, independent of each other. If only the dependent variable is private while the explanatory variable is a public variable, only one condition must be satisfied: noise added to dependent variable is independent of the explanatory variable.

Thus, in contrast to this prevailing wisdom on the disadvantage of differential privacy, our study shows that by declaring the privacy budget used in generating a differentially private dataset, the slope parameters in the original dataset can be retrieved precisely. Thus, our paper is the first to show that differential privacy provides a *precise replication (not approximation as claimed in* [69]*) of the relationships between variables even while preserving the privacy of the dataset*. Our study also contrasts the claim in [69] that "differential privacy provides either very little privacy or very little utility or neither", the thesis in [72] that techniques for privacy preservation have "a noticeable impact of privacy-preservation techniques in predictive performance", and the concerns raised in [5] with respect to the use of differential privacy by the 2020 U.S. Census that "transition to differential privacy has raised a number of questions about the proper balance between privacy and accuracy in official statistics."

Conclusion and future directions

Advances in computing power have enabled unparalleled opportunity for obtaining insights using granular data, especially those on individuals, to guide corporate and public policy. This trend is also accompanied by the increasing importance that society places on individuals' privacy, thereby creating an intrinsic trade-off between the utility of datasets and privacy of individuals that comprise such data. Existing literature highlights this trade-off even for one of the newest concept in privacy—differential privacy. In contrast to such existing literature, our study shows that differential privacy can be employed to *precisely—not approximately—retrieve* the associations in the original dataset provided the noise addition satisfies certain criteria.

Given the promise of differential privacy in preserving the privacy of individuals' data, a follow up to our study could be to study the techniques through which noise can be added to satisfy differential privacy as well as the criteria that are outlined in this study, especially adding noise that is purely random. Another important follow up

study would be to analyze whether the results that we have demonstrated for analysis using ordinary least squares (OLS) regression extend to other analytical techniques, such as those using artificial intelligence and machine learning.

Acknowledgements
I am grateful to Prof. Ashish Goel, Stanford University, for his invaluable guidance. All errors are mine.

Author contributions
As this is a solo-authored paper, I take all the responsibility for the work done.

Funding
No funding was received for this study.

Availability of data and materials
The data used for the analyses are publicly available.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
The author provides their consent for publication of this study.

Competing interests
The author declares that they have no competing interests.

Received: 11 July 2022 Accepted: 2 March 2023
Published online: 14 July 2023

References

1. Na L, Yang C, Lo C-C, Zhao F, Fukuoka Y, Aswani A. Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. JAMA Netw open. 2018;1(8):186040.
2. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS ONE. 2011;6(12):28071.
3. Barth-Jones D. The 're-identification' of governor william weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. Then and Now (July 2012). 2012.
4. Dwork C. Differential privacy. In: International colloquium on automata, languages, and programming. Springer; 2006. p. 1–12.
5. Hawes MB. Implementing differential privacy: seven lessons from the 2020 united states census. 2020.
6. Johnson NM, Near JP, Song DX. Practical differential privacy for sql queries using elastic sensitivity. 2017. arXiv:1706.09479.
7. Pihur V, Korolova A, Liu F, Sankuratripati S, Yung M, Huang D, Zeng R. Differentially-private "draw and discard" machine learning. arXiv preprint. 2018. arXiv:1807.04369.
8. Ross M, Wei W, Ohno-Machado L. Big data and the electronic health record. Yearb Med Inform. 2014;23(01):97–104.
9. Kumar A, Grupcev V, Berrada M, Fogarty JC, Tu Y-C, Zhu X, Pandit SA, Xia Y. Dcms: a data analytics and management system for molecular simulation. J Big Data. 2015;2(1):1–22.
10. Lin C, Song Z, Song H, Zhou Y, Wang Y, Wu G. Differential privacy preserving in big data analytics for connected health. J Med Syst. 2016;40(4):1–9.
11. Moussa M, Demurjian SA. Differential privacy approach for big data privacy in healthcare. In: Privacy and security policies in big data; 2017. p. 191–213.
12. Prasser F, Kohlmayer F, Spengler H, Kuhn KA. A scalable and pragmatic method for the safe sharing of high-quality health data. IEEE J Biomed Health Inform. 2017;22(2):611–22.
13. Yüksel B, Küpçü A, Özkasap Ö. Research issues for privacy and security of electronic health services. Futur Gener Comput Syst. 2017;68:1–13.
14. Zhang R, Xue R, Liu L. Searchable encryption for healthcare clouds: a survey. IEEE Trans Serv Comput. 2017;11(6):978–96.
15. Angeletti F, Chatzigiannakis I, Vitaletti A. Towards an architecture to guarantee both data privacy and utility in the first phases of digital clinical trials. Sensors. 2018;18(12):4175.
16. Swarup S, Braverman V, Arora R, Caragea D, Cragin M, Dy J, Honavar V, Huang H, Locicero R, Singh L, et al. Challenges and opportunities in big data research: Outcomes from the second annual joint pi meeting of the nsf bigdata research program and the nsf big data regional innovation hubs and spokes programs 2018. In: NSF Workshop Reports 2018.
17. Banerjee S, Benlamri R, Bouzefrane S. Optimization of ontology-based clinical pathways and incorporating differential privacy in the healthcare system. Security designs for the cloud, iot, and social networking; 2019. p. 191–205.
18. Harris DR. Leveraging differential privacy in geospatial analyses of standardized healthcare data. In: 2020 IEEE International conference on big Data (Big Data), IEEE; 2020. p. 3119–3122.

19. Bild R, Kuhn KA, Prasser F. Better safe than sorry–implementing reliable health data anonymization. In: Digital personalized health and medicine; 2020. p. 68–72.
20. Hägermalm A, Slavnic S. Differential privacy: an extensive evaluation of open-source tools for ehealth applications. 2021.
21. Huang WA, Kandula A, Wang X. A differential-privacy-based blockchain architecture to secure and store electronic health records. In: 2021 The 3rd International conference on blockchain technology. 2021. p. 189–194.
22. Chong KM, Malip A. Bridging unlinkability and data utility: privacy preserving data publication schemes for healthcare informatics. Comp Commun. 2022. https://doi.org/10.1016/j.comcom.2022.04.032.
23. Truex S, Baracaldo N, Anwar A, Steinke T, Ludwig H, Zhang R, Zhou Y. A hybrid approach to privacy-preserving federated learning. In: Proceedings of the 12th ACM workshop on artificial intelligence and security; 2019. p. 1–11.
24. Tian Z, Zhang R, Hou X, Liu J, Ren K. Federboost: Private federated learning for gbdt. arXiv preprint. 2020. arXiv:2011.02796.
25. Zhang X-Y, Kuenzel S. Differential privacy for deep learning-based online energy disaggregation system. In: 2020 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), IEEE; 2020. p. 904–908.
26. Zhang W, Qiu Y, Bai S, Zhang R, Wei X, Bai X. Fedocr: Communication-efficient federated learning for scene text recognition. arXiv preprint. 2020. arXiv:2007.11462.
27. El Ouadrhiri, A., Abdelhadi, A.: Differential privacy for fair deep learning models. In: 2021 IEEE International Systems Conference (SysCon), IEEE; 2021. p. 1–6.
28. Xu R, Baracaldo N, Zhou Y, Anwar A, Joshi J, Ludwig H. Fedv: Privacy-preserving federated learning over vertically partitioned data. In: Proceedings of the 14th ACM workshop on artificial intelligence and security, 2021; p. 181–192.
29. Liu Z, Zhang R. Privacy preserving collaborative machine learning. EAI Endorsed Trans Secur Saf. 2021;8(28):3.
30. Yu H, Chen Z, Zhang X, Chen X, Zhuang F, Xiong H, Cheng X. Fedhar: Semi-supervised online learning for personalized federated human activity recognition. IEEE transactions on mobile computing. 2021.
31. Zhang R, Song M, Li T, Yu Z, Dai Y, Liu X, Wang G. Democratic learning: hardware/software co-design for lightweight blockchain-secured on-device machine learning. J Syst Archit. 2021;118:102205.
32. Tarun AK, Chundawat VS, Mandal M, Kankanhalli M. Fast yet effective machine unlearning. arXiv preprint. 2021. arXiv:2111.08947.
33. Zhang J, Li C, Robles-Kelly A, Kankanhalli M. Hierarchically fair federated learning. arXiv preprint. 2020. arXiv:2004.10386.
34. Chundawat VS, Tarun AK, Mandal M, Kankanhalli M. Zero-shot machine unlearning. arXiv preprint. 2022. arXiv:2201.05629.
35. Nikolaidis K, Kristiansen S, Plagemann T, Goebel V, Liestøl K, Kankanhalli M, Traaen GM, Overland B, Akre H, Aakerøy L, et al. Learning realistic patterns from visually unrealistic stimuli: generalization and data anonymization. J Artif Intell Res. 2021;72:1163–214.
36. Yu L, Liu L, Pu C, Gursoy ME, Truex S. Differentially private model publishing for deep learning. In: 2019 IEEE symposium on security and privacy (SP), IEEE. 2019. p. 332–349
37. Truex S, Liu L, Chow K-H, Gursoy ME, Wei W. Ldp-fed: Federated learning with local differential privacy. In: Proceedings of the third ACM international workshop on edge systems, analytics and networking; 2020. p. 61–66
38. Wei W, Liu L, Wut Y, Su G, Iyengar A. Gradient-leakage resilient federated learning. In: 2021 IEEE 41st international conference on distributed computing systems (ICDCS), IEEE; 2021. p. 797–807.
39. Karimi L, Palanisamy B, Joshi J. A dynamic privacy aware access control model for location based services. In: 2016 IEEE 2nd international conference on collaboration and internet computing (CIC), IEEE; 2016. p. 554–557.
40. Zhu Y, Wang Y, Liu Q, Liu Y, Zhang P. Wifi fingerprint releasing for indoor localization based on differential privacy. In: 2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC), IEEE; 2017. p. 1–6.
41. Jin L, Li C, Palanisamy B, Joshi J. k-trustee: location injection attack-resilient anonymization for location privacy. Comput Secur. 2018;78:212–30.
42. Wang J, Zhu R, Liu S, Cai Z. Node location privacy protection based on differentially private grids in industrial wireless sensor networks. Sensors. 2018;18(2):410.
43. Yang X, Gao L, Zheng J, Wei W. Location privacy preservation mechanism for location-based service with incomplete location data. IEEE Access. 2020;8:95843–54.
44. Zhang P, Hu C, Chen D, Li H, Li Q. Shiftroute: achieving location privacy for map services on smartphones. IEEE Trans Veh Technol. 2018;67(5):4527–38.
45. Jin B, Zhang Z, Zhao T. Location nearest neighbor query method for social network based on differential privacy. J Comput Appl. 2020;40(8):2340.
46. Kim JW, Edemacu K, Kim JS, Chung YD, Jang B. A survey of differential privacy-based techniques and their applicability to location-based services. Comput Secur. 2021;111:102464.
47. Wen R, Zhang R, Peng K, Wang C. Protecting locations with differential privacy against location-dependent attacks in continuous lbs queries. In: 2021 IEEE 20th International conference on trust, security and privacy in computing and communications (TrustCom), IEEE; 2021. p. 379–386.
48. Dustdar S, Nepal S, Joshi J. Introduction to the special section on advances in internet-based collaborative technologies. New York: ACM; 2019.
49. Sha K, Yang TA, Wei W, Davari S. A survey of edge computing-based designs for iot security. Digit Commun Netw. 2020;6(2):195–202.
50. Husnoo MA, Anwar A, Chakrabortty RK, Doss R, Ryan MJ. Differential privacy for iot-enabled critical infrastructure: a comprehensive survey. IEEE Access. 2021.
51. Jiang B, Li J, Yue G, Song H. Differential privacy for industrial internet of things: Opportunities, applications, and challenges. IEEE Internet Things J. 2021;8(13):10430–51.
52. Hassan MU, Rehmani MH, Chen J. Differential privacy in blockchain technology: a futuristic approach. J Parallel Distrib Comput. 2020;145:50–74.

53. Hassan MU, Rehmani MH, Chen J. Performance evaluation of differential privacy mechanisms in blockchain based smart metering. arXiv preprint. 2020. arXiv:2007.09802.
54. Cao Y, Wei W, Zhou J. Privacy protection data mining algorithm in blockchain based on decision tree classification. In: Web Intelligence, IOS Press; p. 1–10.
55. Hassan MU, Rehmani MH, Chen J. Differential privacy techniques for cyber physical systems: a survey. IEEE Commun Surv Tutor. 2019;22(1):746–89.
56. Olowononi FO, Rawat DB, Liu C. Federated learning with differential privacy for resilient vehicular cyber physical systems. In: 2021 IEEE 18th annual consumer communications and networking conference (CCNC), IEEE; 2021. p. 1–5.
57. Sun X, Yu FR, Zhang P. A survey on cyber-security of connected and autonomous vehicles (cavs). IEEE transactions on intelligent transportation systems. 2021.
58. Lv Z, Chen D, Feng H, Singh AK, Wei W, Lv H. Computational intelligence in security of digital twins big graphic data in cyber-physical systems of smart cities. ACM Trans Manag Inform Syst. 2022. https://doi.org/10.1145/3522760.
59. Xu R, Joshi J, Li C. Nn-emd: Efficiently training neural networks using encrypted multi-sourced datasets. IEEE Transactions on Dependable and Secure Computing. 2021.
60. Zhang J, Sun J, Zhang R, Zhang Y, Hu X. Privacy-preserving social media data outsourcing. In: IEEE INFOCOM 2018-IEEE conference on computer communications. IEEE; 2018. p. 1106–1114.
61. Jiang H, Pei J, Yu D, Yu J, Gong B, Cheng X. Applications of differential privacy in social network analysis: a survey. IEEE transactions on knowledge and data engineering. 2021.
62. Yao X, Zhang R, Zhang Y. Differential privacy-preserving user linkage across online social networks. In: 2021 IEEE/ACM 29th International symposium on quality of service (IWQOS). IEEE; 2021. p. 1–10.
63. Jin X, Zhang R, Chen Y, Li T, Zhang Y. Dpsense: differentially private crowdsourced spectrum sensing. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016. p. 296–307
64. Hu Y, Zhang R. Differentially-private incentive mechanism for crowdsourced radio environment map construction. In: IEEE INFOCOM 2019-IEEE conference on computer communications. IEEE; 2019. p. 1594–1602.
65. Wang D, Ren J, Wang Z, Pang X, Zhang Y, Shen XS. Privacy-preserving streaming truth discovery in crowdsourcing with differential privacy. IEEE transactions on mobile computing. 2021.
66. Ni J, Zhang K, Vasilakos AV. Security and privacy for mobile edge caching: challenges and solutions. IEEE Wirel Commun. 2020;28(3):77–83.
67. Sharma J, Kim D, Lee A, Seo D. On differential privacy-based framework for enhancing user data privacy in mobile edge computing environment. IEEE Access. 2021;9:38107–18.
68. Pejó B, Desfontaines D. Guide to differential privacy modifications: a taxonomy of variants and extensions. Cham: Springer; 2022.
69. Sangeetha S, Sudha Sadasivam G. Privacy of big data: a review. Handbook of big data and iot security. 2019. p. 5–23.
70. Jain P, Gyanchandani M, Khare N. Differential privacy: its technological prescriptive using big data. J Big Data. 2018;5(1):1–24.
71. Stefanowski J, Japkowicz N. Final remarks on big data analysis and its impact on society and science. In: Big data analysis: new algorithms for a new society. Springer; 2016. p. 305–329.
72. Carvalho T, Moniz N. The compromise of data privacy in predictive performance. In: International symposium on intelligent data analysis. Springer; 2021. p. 426–438.
73. Zhang L, Niu D, Li Y, Zhang Z. A survey on privacy preserving association rule mining. In: 2018 5th International conference on information science and control engineering (ICISCE). IEEE; 2018. p. 93–97.
74. Ahluwalia MV, Gangopadhyay A, Chen Z, Yesha Y. Target-based, privacy preserving, and incremental association rule mining. IEEE Transact Serv Comput. 2015;10(4):633–45.
75. Wang J, Liu S, Li Y. A review of differential privacy in individual data release. Int J Distrib Sensor Netw. 2015;11(10):259682.
76. Angrist JD, Pischke J. Mostly harmless econometrics: an empiricist's companion. Princeton: Princeton University Press; 2009.
77. Wooldridge J, Imbens G. Difference-in-differences estimation. Lecture notes. 10. 2007.

Publisher's Note