Privacy-preserving data sharing via probabilistic modelling

Joonas Jälkö *
Department of Computer Science
Aalto University
Espoo, Finland
joonas.jalko@aalto.fi

Antti Honkela
Department of Computer Science
University of Helsinki
Helsinki, Finland

Samuel Kaski
Department of Computer Science
Aalto University
Espoo, Finland

Abstract

Differential privacy allows quantifying privacy loss from computations using sensitive personal data. This loss grows with the number of accesses to the data, making it hard to open the use of such data while respecting privacy. Instead of accessing the data multiple times, we propose a method of fitting a probabilistic model on the data using privacy preserving modelling techniques. From this probabilistic model we sample a new synthetic dataset that may be subjected to unlimited amount of future analysis, without affecting the privacy guarantees. We demonstrate empirically that similar statistical discoveries can be made from the synthetic as the original data. We expect the method to have broad use in sharing anonymized versions of key data sets for research.

1 Introduction

Releasing datasets would be beneficial for the research community. However in general this is not possible, due to the sensitive nature of information contained in for instance medical and many other datasets. Recent advances in privacy-preserving machine learning have opened a possibility for an alternative way towards opening the use of data: to learn from the sensitive dataset without violating the anonymity of the individuals in the dataset.

Differential privacy (DP) [9] gives a statistical measure of privacy and anonymity. It provides strict controls on the level to which an individual can be identified from the result of an algorithm operating on personal data.

DP is widely used, but usually only to answer a specific question rather than releasing the data. While we nowadays have powerful privacy-preserving tools for variety of machine learning methods, for example the widely popular differentially private SGD [1], there is no simple way to bound the privacy loss of answering arbitrary queries under DP. Another obstacle on the way towards widespread use is that the data will be queried multiple times, the privacy risk will accumulate.

Releasing data under privacy guarantees would enable unlimited further operations on the dataset. Recently, many privacy-preserving synthetic data release techniques based on deep learning have been proposed [3, 17, 2, 4, 16, 19]. However, using Bayesian approach for generating data has not been widely studied. In the past there have been some suggestions for privacy-preserving data release methods for specific probabilistic models [6, 18].

In this paper we formulate the general principle of *Bayesian DP data sharing* and demonstrate how to successfully apply it in two different model families. With empirical experiments we show that very similar discoveries can be made from the synthetic data as we would from the original data.

2 Differentially private data sharing

We recall some basic concepts of differential privacy.

Definition 1 (Differential privacy)**.** *A randomized algorithm $\mathcal{A} : \mathcal{X}^N \to \mathcal{I}$ satisfies (ϵ, δ) differential privacy, if for all adjacent datasets $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^N$ and for all measurable $I \subset \mathcal{I}$ it holds that*

$$\Pr(\mathcal{A}(\mathbf{x}) \in I) \leq e^{\epsilon} \Pr(\mathcal{A}(\mathbf{x}') \in I) + \delta. \tag{1}$$

DP has many desirable properties such as composability of privacy quarantees, which allows quantifying privacy parameters of multiple applications of DP algorithms. In the most basic form, the privacy cost of the compositional query will be $(\sum_i \epsilon_i, \sum_i \delta_i)$ [8], where (ϵ_i, δ_i) is a cost of the ith query. Another important property of DP is invariance to post-processing [10]. This means that the privacy guarantees of a DP results cannot be degraded by further manipulations of the result. Thus we can use results of DP algorithms to answer future queries under the same privacy guarantees.

To formulate our framework of Bayesian DP data sharing, lets consider a dataset $\mathbf{X}$ and a probabilistic model $\mathcal{M}(\mathbf{X}, \mathbf{Z})$ with latent variables $\mathbf{Z}$. Our aim is to release a new synthetic dataset $\tilde{\mathbf{X}}$ by learning a data-generating model based on the original data. We will choose posterior predictive distribution $p(\tilde{\mathbf{X}} \,|\, \mathbf{X})$ as our generative model

$$p(\tilde{\mathbf{X}} \,|\, \mathbf{X}) = \int_{\text{Supp}(\mathbf{Z})} p(\tilde{\mathbf{X}} \,|\, \mathbf{Z}) p(\mathbf{Z} \,|\, \mathbf{X}) \mathrm{d}\mathbf{Z}. \tag{2}$$

We sample from posterior predictive distribution, by first drawing $\tilde{\mathbf{Z}}$ from the posterior distribution $p(\mathbf{Z} \,|\, \mathbf{X})$ and then draw new data sample $\tilde{\mathbf{x}}$ from $\mathcal{M}$ conditioned on $\tilde{\mathbf{Z}}$.

As we access the data only through the posterior distributions of the latent variables, it suffices to learn these distributions under DP. Sampling from these posteriors can be considered as post-processing, and thus no further effects on the privacy. Recently proposed privacy preserving Bayesian inference methods [12, 15, 13, 11] aim to provide a solution for DP posterior inference.

The data sharing method is flexible and we can use any probabilistic model, that can be trained under differential privacy, as the generator. In this paper, we study two different model families, Bayesian networks and mixture models as the probabilistic model.

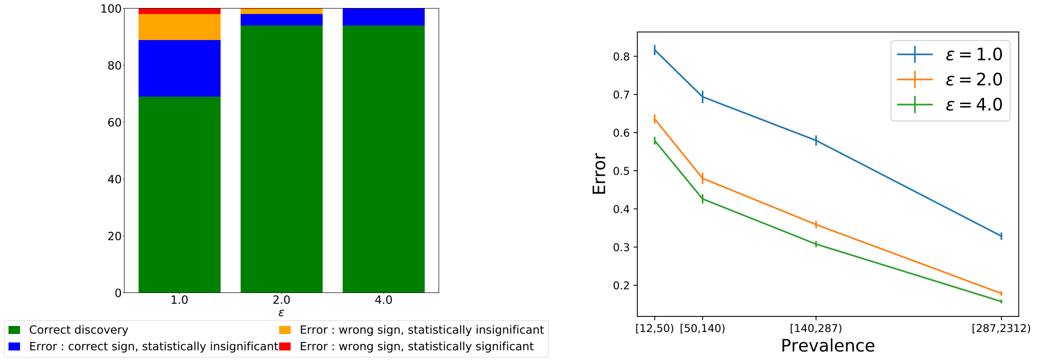
3 Case studies

3.1 Making statistical discoveries from the synthetic data

To test whether the same discoveries can be done from the synthetic as from the original data set, we generated a synthetic data set based on an epidemiological set [5], using a general-purpose generative model family (mixture model). The original data comprised of 208 148 females and 226 372 males.

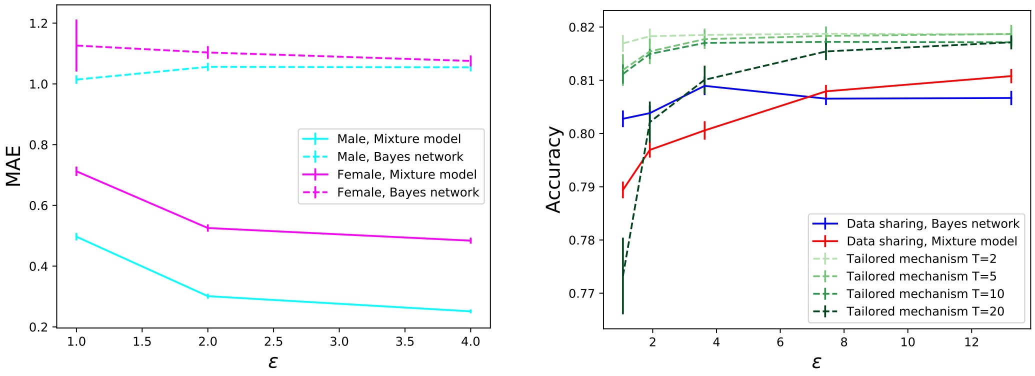
The data have previously been used to study the effect of diabetes medication to alcohol related deaths (ARD) using a Poisson regression model [14]. We fit a similar Poisson regression model to the synthetic data and compared. For males, the results is that we can make the same statistical discovery from the synthetic data : the diabetics have a higher mortality rate than the non-diabetics (Figure 1a). Even with a reasonable level of privacy ($\epsilon = 2.0$), it is very likely that we make the correct statistical discovery from the synthetic data.

Since DP guarantees indistinguishability among individuals in the dataset, differentially private algorithms are bound to lose some of the rare characteristics of the data. To assess this, we split the regression coefficients, both male and female, into four equal sized bins of corresponding prevalence and computed the mean absolute error between original and synthetic coefficients within these bins. Figure 1b shows that the regression coefficients with higher prevalence are more accurately discovered from the synthetic data. By prevalence, we refer to the number of individuals with follow-up ending at an incidence.



(a) **ARD study**. The correct statistical discoveries can be made with high probability from the synthetic data. Bars show the percentage of correct discoveries in 100 runs and the different types of errors as a function of the privacy level epsilon.

(b) **ARD study**. Accuracy of synthetic regression coefficients improves as the number of relevant examples grows. Mean absolute error between original and synthetic coefficients of coefficients within a prevalence bin. Prevalence denotes the min and max number of relevant examples within a bin. Average result over 100 independent runs of the algorithm. Error bars : standard error of mean.



(a) **ARD study**. Mixture models preserve regression coefficients better than the Bayes network. Average over 100 runs. Error bars : standard error of mean.

(b) **Adult study**. Both data sharing methods beat the tailored mechanism in the high privacy region as the number of anticipated queries grows. Average of 10 runs. Errorbars : standard error of mean.

3.2 Choice of probabilistic model

We test the of effect of probabilistic model by comparing results obtained from synthetic data of two different probabilistic models : mixture model and private Bayes networks [18]. We evaluate the performance of both models with two datasets : the epidemiological data discussed in Section 3.1 and the publicly available Adult dataset [7].

In the ARD study discussed in Section 3.1, the mixture models perform better than the Bayesian network approach. Figure 2a shows the accuracy of both probabilistic models in terms of mean absolute error between regression coefficients obtained from original and synthetic data.

We also compared the two probabilistic models in a classification task using the Adult dataset. After learning the generative model, we used the synthetic data obtained from the generative model to train a logistic regression classifier, to predict whether the individuals annual income exceeds \$50 000. We used a separate test set to test the performance of the method. Figure 2b illustrates that in this example, the Bayes network outperforms the mixture model in terms of classification accuracy in the strict privacy region.

3.3 Performance against tailored mechanism

We compared the synthetic data release method in Adult example against a private logistic regression classifier in a case where the data holder would split the privacy budget uniformly among for T queries. Figure 2b shows that as the privacy budget increases, the tailored mechanism outperforms data release mechanism with both Bayes network and mixture model. However, in the strict privacy region, data-sharing method with both probabilistic models performs better than the tailored mechanism if data holder would prepare for 20 queries.

4 Conclusions

We have presented a privacy-preserving data sharing mechanism that is applicable for arbitrary datasets. Our data sharing method allows unlimited number of arbitrary tasks to be performed on the synthetic data with no further privacy considerations. This is especially beneficial for tasks for which there is no existing privacy preserving counterpart. The method works better with large datasets. This is because of the nature of DP, it is easier to mask the contribution of one element of the dataset when the data size is large.

Choosing the correct probabilistic model could possibly have a huge impact on the performance of the method. For example with the epidemiological study dataset, the follow-up for each individual ended either on date of death or on the end of study. This implies that the follow-up date, start date of the follow-up and the occurrence of death have a dependency structure, thus it makes sense to include this kind of structure directly to the model rather than waste the expressiveness of the probabilistic model to learn it. Also as we saw in the Adult example, the Bayesian networks performed better than the mixture model in the high privacy region. With relatively small dataset such as Adult, one should use probabilistic model such as Bayesian networks to describe it.

References

[1] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. 2016. arXiv:1607.00133 [stat.ML].

[2] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 510–526. Springer, 2018.

[3] Gergely Acs, Luca Melis, Claude Castelluccia, and Emiliano De Cristofaro. Differentially private mixture of generative neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[4] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *bioRxiv*, page 159756, 2017.

[5] Anna But, Marie L De Bruin, Marloes T Bazelier, Vidar Hjellvik, Morten Andersen, Anssi Auvinen, Jakob Starup-Linde, Marjanka K Schmidt, Kari Furu, Frank de Vries, et al. Cancer risk among insulin users: comparing analogues with human insulin in the CARING five-country cohort study. *Diabetologia*, 60(9):1691–1703, 2017.

[6] Rui Chen, Qian Xiao, Yu Zhang, and Jianliang Xu. Differentially private high-dimensional data publication via sampling-based inference. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 129–138, New York, NY, USA, 2015. ACM.

[7] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[8] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC*, volume 9, pages 371–380, 2009.

[9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC 2006*. 2006.

[10] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014.

[11] Mikko A Heikkilä, Joonas Jälkö, Onur Dikmen, and Antti Honkela. Differentially private markov chain monte carlo. *arXiv preprint arXiv:1901.10275*, 2019.

[12] Joonas Jälkö, Onur Dikmen, and Antti Honkela. Differentially private variational inference for non-conjugate models. In *Uncertainty in Artificial Intelligence 2017 Proceedings of the 33rd Conference, UAI 2017*. The Association for Uncertainty in Artificial Intelligence, 2017.

[13] Bai Li, Changyou Chen, Hao Liu, and Lawrence Carin. On connecting stochastic gradient MCMC and differential privacy. *arXiv preprint arXiv:1712.09097*, 2017.

[14] Leo Niskanen, Timo Partonen, Anssi Auvinen, and Jari Haukka. Excess mortality in Finnish diabetic subjects due to alcohol, accidents and suicide: a nationwide study. *European Journal of Endocrinology*, 1(aop), 2018.

[15] Mijung Park, James Foulds, Kamalika Chaudhuri, and Max Welling. Variational Bayes in private settings (VIPS). *arXiv preprint arXiv:1611.00340*, 2016.

[16] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[17] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations (ICLR 2019)*, 2019.

[18] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. PrivBayes: Private data release via Bayesian networks. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1423–1434, New York, NY, USA, 2014. ACM.

[19] Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model. *arXiv preprint arXiv:1801.01594*, 2018.