PERFECTLY FAIR AND DIFFERENTIALLY PRIVATE SELECTION USING THE LAPLACE MECHANISM *

Mina Samizadeh, Mohammad Mahdi Khalili
Department of Information and Computer Sciences
University of Delaware
Newark, DE 19711, USA
{minasmz, khalili}@udel.edu

ABSTRACT

Supervised machine learning is widely used for selection problems where a limited number of individuals are selected from an applicant pool. Many real-world problems such as hiring and university admission can be modeled as selection problems. Machine learning models often suffer from a bias against certain demographic groups due to pre-existing biases in training datasets. In addition to the (un)fairness issue, privacy concerns arise when models are trained on sensitive personal information. In this work, we are interested in understanding the effect of privacy-preserving mechanisms on fairness in selection problems. In particular, we consider a scenario where a machine learning model is used to generate a qualification score and adopt the *Laplace mechanism* to achieve the ϵ-differentially privacy. In this scenario, we identify conditions under which the scores generated by the *Laplace mechanism* lead to perfect fairness in selection problems.

1 INTRODUCTION

Supervised machine learning models are powerful tools in decision-making problems. In particular, they can be used in selection problems where a limited number of individuals are selected from an applicant pool. Many real-world problems such as university admission, resource allocation, and hiring can be regarded as selection problems. However, machine learning models trained on real-world data capture the pre-existing biases in training datasets Lee et al. (2019); De-Arteaga et al. (2019); Dressel & Farid (2018). Various fairness notions (e.g., equal opportunity, statistical parity) have been introduced in the literature to address fairness issues. For instance, the equal opportunity notion implies that the true positive rates should be the same across different demographic groups Hardt et al. (2016), and the statistical parity fairness notion requires the output of a machine learning model to be independent of sensitive attributes Dwork et al. (2012). In addition to fairness issues, privacy concerns may arise when the personal information of individuals is involved in the training or decision-making process. Among various privacy notions, differential privacy has gained much attention Dwork et al. (2006a) Dwork et al. (2006b) and has been widely used in practice. Several mechanisms have been introduced to ensure differential privacy.

In this work, we adopt the *laplace mechanism* Dwork et al. (2006b) to ensure a privacy guarantee in selection problems. In particular, we use the *laplace mechanism* as a post-processing tool to generate and assign a qualification score to each applicant. We show that the *laplace mechanism* is able to improve fairness in selection problems if the selection rule is to select an individual with the highest qualification score. In particular, we find conditions under which perfect fairness is achievable using the *laplace mechanism*. On the other hand, without the *laplace mechanism*, selecting an applicant with the highest qualification score does not ensure fairness. It also compromises the privacy of each applicant.

In our previous work Khalili et al. (2020), we observed and proved that fairness and privacy are compatible under the exponential mechanism McSherry & Talwar (2007). In this work, we show that

*This article is an extended abstract for the work accepted in the ICLR workshop on Socially Responsible Machine Learning (SRML)

the *laplace mechanism* gives us more degrees of freedom compared to the exponential mechanism. More specifically, we show that the *laplace mechanism* enables us to add different noise levels to the qualification scores based on the sensitive attributes. As a result, we are able to satisfy perfect fairness under milder conditions compared to those found in Khalili et al. (2020).

REFERENCES

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pp. 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006b.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Mohammad Mahdi Khalili, Xueru Zhang, Mahed Abroshan, and Somayeh Sojoudi. Improving fairness and privacy in selection problems. *arXiv preprint arXiv:2012.03812*, 2020.

Paulyne Lee, Maxine Le Saux, Rebecca Siegel, Monika Goyal, Chen Chen, Yan Ma, and Andrew C Meltzer. Racial and ethnic disparities in the management of acute pain in us emergency departments: meta-analysis and systematic review. *The American journal of emergency medicine*, 37 (9):1770–1777, 2019.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103. IEEE, 2007.