On the Duality of Privacy and Fairness

Mário S. Alvim, Natasha Fernandes, Bruno D Nogueira, Catuscia Palamidessi, Thiago V A Silva

On the Duality of Privacy and Fairness (Extended Abstract)

Mário S. Alvim[1], Natasha Fernandes[2], Bruno D. Nogueira[1], Catuscia Palamidessi[3], and Thiago V.A. Silva[1]

[1] *UFMG, Brazil*
[2] *Macquarie University, Australia*
[3] *Inria and LIX, École Polytechnique, France*

Abstract

When a machine learning model operates over data about individuals, there are two common concerns. On one hand, if the model's output (i.e., its prediction) allows for information inferences about an individual's sensitive attributes, we have a *privacy* issue. On the other hand, if the individual's sensitive attributes can unduly influence the model's output, we have a *fairness* issue. Recently, the interplay between these two concerns has gathered growing attention both in the scientific community and in society as a whole. In this work, we extend the framework of *quantitative information flow* to formally capture fairness and privacy as duals of each other, and give first steps toward a novel characterization of their relationship.

1 Introduction

The increasing ubiquity of machine learning (ML) has brought not only significant benefits to modern society, but also some pressing concerns. One such concern is that ML models can be *unfair*, in the sense that they may unduly use individuals' sensitive attributes (such as race or gender) to make decisions, thereby creating or reinforcing biases against certain groups. A second concern is that ML models may *not be private*, i.e., they may improperly leak individuals' sensitive attributes in the course of their execution. These issues, and their interrelationship, have gained increased attention [1,2,8,9,12,19,22].

In this work, we employ the framework of *quantitative information flow (QIF)* [4] to investigate the interplay between fairness and privacy in ML. QIF has been successfully applied to a wide range of privacy-related problems [5, 10, 11, 15–17], but, to the best of our knowledge, not yet to fairness. Here we report first steps in modeling fairness in QIF as a dual concept to privacy, and conjecture a formal trade-off between them.

2 Preliminaries

In this section, we briefly review some key concepts from QIF, ML, privacy, and fairness needed for this work.

Quantitative information flow (QIF) and privacy. QIF is concerned with measuring the amount of information that a system leaks about its (secret) inputs through its (observable) execution behaviour to an *adversary*, which is an entity trying to infer sensitive information.

Before the system is run, the adversary has some *a priori knowledge* about the secret values, represented by a *prior distribution* (or simply a *prior*) $\pi{:}\mathbb{D}\mathcal{X}$ on the set $\mathcal{X}$ of possible values for the secret. (Given a set $\mathcal{S}$, we denote by $\mathbb{D}\mathcal{S}$ the set of all probability distributions over $\mathcal{S}$.) We let π_x denote the prior probability of secret value $x{\in}\mathcal{X}$. We can measure the *prior vulnerability* of prior π as the adversary's probability of guessing the secret value correctly in one try. This is formalised as *Bayes vulnerability*, defined as $V(\pi){=}\max_{x\in\mathcal{X}}\pi_x$.

A system is modeled as an *(information-theoretic) channel* $C{:}\mathcal{X}{\to}\mathbb{D}\mathcal{Y}$ mapping (in a possibly probabilistically way) secret inputs from $\mathcal{X}$ to observable behaviours from a set $\mathcal{Y}$. We let $C_{x,y}$ denote the probability of channel C outputting $y{\in}\mathcal{Y}$ when its input is $x{\in}\mathcal{X}$. We assume that the adversary knows how the system works (i.e., the channel C) and, from that and the prior π, can derive a joint distribution $\pi{\triangleright}C{:}\mathbb{D}(\mathcal{X}{\times}\mathcal{Y})$ on inputs and outputs of a system. More precisely, $(\pi{\triangleright}C)_{x,y}{=}\pi_x C_{x,y}$ for all $x{\in}\mathcal{X}$, $y{\in}\mathcal{Y}$. This joint represents the adversary's *a posteriori knowledge* about the secret (i.e., after the system is run), and from it we can compute the *posterior Bayes vulnerability* of the secret as $V[\pi{\triangleright}C]{=}\sum_{y\in\mathcal{Y}} max_{x\in\mathcal{X}}(\pi{\triangleright}C)_{x,y}$. This value represents the adversary's expected probability of guessing the secret correctly in one try after the system is run.

The *Bayes (information) leakage* caused by system C under prior π can be quantified as the ratio by which the system's execution increases the secret's Bayes vulnerability, and it is given by $\mathcal{L}(\pi, C){=}{}^{V[\pi\triangleright C]}\!/_{V(\pi)}$. This leakage from secret inputs to observable outputs is a measure of the system's *privacy*.

Machine learning and fairness. In a *classification problem*, a ML model receives as input a vector of *features* for an individual and outputs a prediction for another feature of this individual. As usual in the fairness literature, we consider *binary classification problems*, in which the model's input is a single, binary sensitive attribute, and its output is a single, binary predicted class. More precisely, the input is a sensitive attribute x taking values s_0/s_1 indicating, resp., whether or not the individual belongs to a protected group (e.g., an ethnic minority/majority), and the output y is a class prediction taking values $+/-$ indicating, resp., a desirable or non-desirable outcome (e.g., acceptance/rejection to a university application). The unfairness of a model is often measured as *statistical disparity* [7, 14, 20, 21, 23], i.e., the dependence level of the classification outcome on the sensitive attribute.

3 Fairness as reverse leakage in QIF

The crucial idea of our modeling of fairness in QIF as a dual of privacy is the novel concept of "reverse leakage".

Recall from the previous section that, given a prior π and a channel C, the corresponding leakage $\mathcal{L}(\pi, C)$ is a measure of how much information the adversary gains about the channel's sensitive input after observing the channel's output. For instance, consider an ML scenario in which the sensitive attribute x represents an individual's health status (e.g., the presence or absence of a certain disability), and the output of the classifier indicates whether or not that individual should be admitted to a particular university. If by observing the output of the classifier (e.g., rejection of the individual's application) the adversary can gain information about the individual's health status (e.g., infer that the individual is likely disabled), then intuitively we have a breach of the individual's privacy. This notion of leakage, which we call *direct*, has been extensively used in the literature as a measure of privacy as resistance to inferences about sensitive values [3, 5, 6, 15].

Here we introduce the novel notion of "reverse leakage" in QIF, which is the dual of direct leakage. Reverse leakage measures the amount of information about the channel's output that an adversary can infer (or predict) if she knows the channel's sensitive input. As an example, consider again the scenario of the ML classifier for university admissions introduced above. In this case, a suitable notion of "reverse leakage" would quantify how well an adversary can predict the classifier's outcome (admission/rejection) given knowledge of the sensitive attribute (disability/non-disability) given as input. Intuitively, the greater the reverse leakage, the more the sensitive attribute influences the classification outcome and, consequently, the less fair the ML model would be. This is the same principle *statistical disparity* and its variants [7, 14, 20, 21, 23]. It is natural, then, to propose "reverse leakage" as a measure of "unfairness" of a classifier given a prior.

Definition 1 (Reverse-prior, –vulnerabilities, and –leakage)**.** *Let $\pi{:}\mathbb{D}\mathcal{X}$ be a prior and $C{:}\mathcal{X}{\to}\mathbb{D}\mathcal{Y}$ be a channel. Then:*

(a) The corresponding reverse prior $\rho^{\pi,C}{:}\mathbb{D}\mathcal{Y}$ *is a distribution on the channel's outputs obtained by marginalizing the joint $\pi{\triangleright}C$ on $\mathcal{Y}$: $\rho^{\pi,C}_y = \sum_{x\in\mathcal{X}}(\pi{\triangleright}C)_{x,y}$ for all $y{\in}\mathcal{Y}$.*

(b) The corresponding reverse prior Bayes vulnerability *is given by $V(\rho^{\pi,C}) = \max_{y\in\mathcal{Y}} \rho^{\pi,C}_y$, and it represents the adversary's probability of guessing the system's output correctly without access to the system's input.*

(c) The corresponding reverse posterior Bayes vulnerability *is given by $V^{rev}[\pi{\triangleright}C] = \sum_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}}(\pi{\triangleright}C)_{x,y}$, and it represents the adversary's probability of guessing the system's output correctly after having access to the system's input.*

(d) The corresponding reverse Bayes leakage *is given by $\mathcal{L}^{rev}(\pi, C) = {}^{V^{rev}[\pi{\triangleright}C]}/_{V(\rho^{\pi,C})}$, and it represents the adversary's gain of prediction power about the channel's output.*

Now we are ready to define unfairness in QIF.

Definition 2 (Quantification of unfairness)**.** *Let $C{:}\mathcal{X}{\to}\mathbb{D}\mathcal{Y}$ be a ML model (probabilistically) mapping sensitive attributes in $\mathcal{X}$ to a classification value in $\mathcal{Y}$. Let $\pi{:}\mathbb{D}\mathcal{X}$ be a prior on the model's inputs. Then the* quantification of the unfairness of C under prior π *is given by the reverse leakage $\mathcal{L}^{rev}(\pi, C)$.*

4 On the fairness-privacy relation

The formalization of fairness as the dual of privacy in QIF allows for the investigation of the relationship between these concepts given a classifier and a prior.

Indeed, we applied our model to compute the fairness and privacy levels for all possible joints $[\pi{\triangleright}C]$ on sensitive input attributes and observable outputs of binary classifiers, up to a resolution of 2^{-9} units of probability mass. Figure 1 shows the results, where each blue point represents a joint, and its position in the graph represents its level of unprivacy (horizontal axis) and unfairness (vertical axis), measured as, respectively, direct $\mathcal{L}(\pi, C)$ and reverse $\mathcal{L}^{rev}(\pi, C)$ Bayes leakage. Higher values of leakage (closer to 2) represent more unfair/unprivate joints, and lower values (closer to 1) represent the opposite.

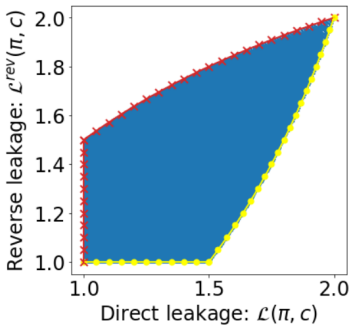

Figure 1. Direct and reverse leakages in binary classifiers.

Note that, given the duality of our formulation of fairness and privacy, the graph is symmetric around the line $y{=}x$. The blue region is the feasible region for direct– and reverse Bayes-leakage. Note also that not all combinations are possible, so it is interesting to identify the graph's Pareto curves, i.e., the maximum level of fairness that can be achieved for a given level of privacy (the yellow "●"-curve), and vice-versa (the red "×"-curve). Indeed, we have the following conjecture for the exact characterization of these Pareto curves.

Conjecture 3 (Characterization of the Pareto curves of fairness-privacy)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be binary sets. Then, for every prior $\pi{:}\mathbb{D}\mathcal{X}$ and channel $C{:}\mathcal{X}{\to}\mathbb{D}\mathcal{Y}$, the relationship ${}^{\mathcal{L}^{rev}(\pi{\triangleright}C)}/_{(3-\mathcal{L}^{rev}(\pi{\triangleright}C))} \leq \mathcal{L}(\pi{\triangleright}C) \leq {}^{(3\cdot\mathcal{L}^{rev}(\pi{\triangleright}C))}/_{(\mathcal{L}^{rev}(\pi{\triangleright}C)+1)}$ holds, and equality can be achieved on both ends.*

5 Conclusion

In this work we introduced a formulation of fairness as the dual of privacy in the QIF framework. This notion measures the dependence of the classification outcome on the sensitive attribute, and it is akin to *statistical disparity* and its variants. From that, we conjectured the Pareto curves describing the optimal trade-off between privacy and fairness for joints arising from binary classifiers. As future work we want to generalise our analysis from Bayes vulnerability to the full framework of g-vulnerabilities provided by QIF, and also extend the notion to capture scenarios in which the disparity is justified by legitimate controlling factors, following the principle that motivates the notion of *conditional statistical parity* [13, 18]. The long-term plan is to use our formalisation to investigate the behaviour of common ML algorithms on real-life datasets and devise methods to enhance fairness in the predictions.

Acknowledgements. Mário S. Alvim, Bruno D. Nogueira and Thiago V. A. Silva were partially funded by CNPq, CAPES, and FAPEMIG. Catuscia Palamidessi was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme. Grant agreement № 835294.

References

[1] Sushant Agarwal. Trade-offs between fairness, interpretability, and privacy in machine learning. Master's thesis, University of Waterloo, 2020.

[2] Mohammad Al-Rubaie and J Morris Chang. Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy*, 17(2):49–58, 2019.

[3] Mário S. Alvim, Miguel E. Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. On the information leakage of differentially-private mechanisms. *Journal of Computer Security*, 23(4):427–469, 2015.

[4] Mário S. Alvim, Konstantinos Chatzikokolakis, Annabelle McIver, Carroll Morgan, Catuscia Palamidessi, and Geoffrey Smith. *The Science of Quantitative Information Flow*. Springer, 2020.

[5] Mário S. Alvim, Natasha Fernandes, Annabelle McIver, Carroll Morgan, and Gabriel H. Nunes. Flexible and scalable privacy assessment for very large datasets, with an application to official governmental microdata. *Proc. of the 22nd Privacy Enhancing Technologies Symposium (PoPETS 2022)*, 2022(4):378–399, 2022.

[6] Mário S. Alvim, Natasha Fernandes, Annabelle McIver, and Gabriel H. Nunes. On privacy and accuracy in data releases (invited paper). In *Proc. of the 31st International Conference on Concurrency Theory (CONCUR 2020)*, volume 171 of *LIPIcs*, pages 1:1–1:18, 2020.

[7] Solon Barocas and Moritz Hardt. NIPS 2017 Tutorial on Fairness in Machine Learning, 2017.

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. of the 2nd Conference on Fairness, Accountability and Transparency (FAT 2019)*, pages 77–91. PMLR, 2018.

[9] Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. In *Proc. of the 6th IEEE European Symposium on Security and Privacy (EuroS&P 2021)*, pages 292–303. IEEE, 2021.

[10] Konstantinos Chatzikokolakis, Natasha Fernandes, and Catuscia Palamidessi. Comparing systems: Max-case refinement orders and application to differential privacy. In *Proc. of the 32nd IEEE Computer Security Foundations Symposium (CSF 2019)*, pages 442–457, 2019.

[11] Giovanni Cherubin, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. F-BLEAU: fast black-box leakage estimation. In *Proc. of the 40th IEEE Symposium on Security and Privacy (S&P 2019)*, pages 835–852, 2019.

[12] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

[13] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. In *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017)*, page 797–806, 2017.

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proc. of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012)*, page 214–226, 2012.

[15] Danilo Fabrino Favato, Gabriel Coutinho, Mário S. Alvim, and Natasha Fernandes. A novel reconstruction attack on foreign-trade official statistics, with a brazilian case study. *Proc. of the 22nd Privacy Enhancing Technologies Symposium (PoPETS 2022)*, 2022(4):608–625, 2022.

[16] Natasha Fernandes, Mark Dras, and Annabelle McIver. Processing text for privacy: an information flow perspective. In *Proc. of the 22nd International Symposium on Formal Methods (FM 2018)*, pages 3–21. Springer, 2018.

[17] Mireya Jurado, Catuscia Palamidessi, and Geoffrey Smith. A formal information-theoretic leakage analysis of order-revealing encryption. In *Proc. of the 34th IEEE Computer Security Foundations Symposium (CSF 2021)*, pages 1–16. IEEE, 2021.

[18] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35:613–644, 2013.

[19] Mohammad Mahdi Khalili, Xueru Zhang, Mahed Abroshan, and Somayeh Sojoudi. Improving fairness and privacy in selection problems. *Proc. of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)*, 35:8092–8100, 05 2021.

[20] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Proc. of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, page 4069–4079, 2017.

[21] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193 – 1216, 2017.

[22] Tianqing Zhu, Dayong Ye, Wei Wang, Wanlei Zhou, and Philip S. Yu. More than privacy: Applying differential privacy in key areas of artificial intelligence. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 34(6):2824–2843, 2020.

[23] Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *CoRR*, abs/1505.05723, 2015.