Research Project - Masters - PPGCC - DCC - UFMG

Artur Gaspar da Silva

This document presents the research project for the candidate Artur Gaspar da Silva, for PPGCC - UFMG. The research area of the project is included in Computer Science Theory, more specifically in Quantitative Information Flow and Privacy. The student aims to be advised by professor Mário Sérgio Alvim, researcher at the T-Rex (Theory Expertise) laboratory.

1 Introduction

Recent resarch[12][1][7][9] indicates numerous tensions and synergies between many concepts that surround the Machine Learning literature, including Fairness, Privacy, Accuracy and Interpretability. For instance, there is an inherent tradeoff between Fairness and Accuracy such that, depending on the data distribution, it might be impossible to develop a model that achieves acceptable values for both fairness and accuracy, if we considr some reasonable fairness metrics[20]. Also, there has been some work on introducing Causality concepts into the discussion, for instance, to develop better fairness metrics[18]. It has also been suggested to use interpretable models (as explanations to more complex models) for auditing systems and checking if they are fair, although this might lead to problems[24]. This area of research is especially relevant nowadays, given the importance that Machine Learning and Artificial Intelligence systems have: we now have computational systems that are part of processes of making decisions with big impacts on people's lives, for instance, recidivism prediction[6], loan approvals[23], hiring decisions[15], and others.

The main goal is to identify the theoretical relations between fairness, privacy and Quantitative Information Flow, which is aligned with research areas at the T-Rex (Theory Expertise) laboratory. More specifically, the aim is to model Differential Privacy and Local Differential Privacy with the Quantitative Information Flow framework, and prove new theoretical results that establish this relation. Afterwards, we aim to explore the relationships between privacy and fairness metrics, with focus on Equal Opportunity Difference, Statistical Disparity and Conditional Statistical Disparity. Finally, causalily concepts (based on Judea Pearl's work[19]) will be explored to provide better mathematical understanding of fairness and privacy in machine learning systems.

2 Theoretical Reference

Causality refers to the study of causal relationships between variables, and how to model and infer causal relationships from the combination of domain knowledge and data[19]. This area of research has matured a lot in the last 50 years, with many different approaches still being developed. Fairness in Machine Learning is concerned with measuring how unfair

the results provided by Machine Learning models are to certain groups or individuals[17], and improving how fair the models are[14]. There are tensions between different fairness measures[13][2]. Privacy is concerned with quantifying how much sensitive information leaks about individuals and methods to avoid this information leakage. In Machine Learning settings, the data collection might be hard for information that is considered very sensitive (for instance, whether or not a person regularly uses illegal drugs) and approaches such as Differential Privacy[8] might improve trust in the data collection. Also, the model itself might allow the identification of individuals and sensitive features, which is not desirable[16]. Accuracy is a metric of how many mistakes the Machine Learning model makes, and there are trade-offs between Accuracy and the other concepts presented[12][20][7]. The area of Interpretability focus on developing Machine Learning models that have human-comprehensible decisions (either directly or to explain the decisions of more complex models), which might be useful when developing these models[22] and also to help experts with domain knowledge decide when to trust the results presented by the models[21]. Quantitative Information Flow is a general theoretical framework for measuring amounts of information, with a focus on privacy applications but, in principle, a broader scope[5].

previous work extensively explored the relationships between Fairness, Interpretability and Privacy[12]. Other works focus on: relationships between Privacy and Fairness[9], the relationship between Privacy, Fariness and Accuracy[7], the feasibility regions of Accuracy and Fairness metrics[20][1], and Causality-Aware fairness metrics[18]. There are also explorations of the relations between Quantitative Information Flow and Fairness[4].

More specifically to the relation between Differential Privacy and Quantitative Inforamtion Flow, there are important results in the literature. There are works discussing the relations between differential privacy and g-vulnerability, including bounds on g-leakage as a function of the ϵ parameter of differential privacy, and the fact that there is no bound on differential privacy as a function of the g-vulnerability [3]. Also, we have recent work [10] discussing how the ϵ parameter of Differential Privacy is related to max-case g-vulnerability: e^{ϵ} is exactly the multiplicative max-case channel capacity under a fixed prior. This work also discusses many other theoretical results relating g-vulnerability notions with differential privacy. Finally, recent contributions show that it is possible to model the ϵ parameter of local differential privacy with the Quantitative Information Flow framework [11], and thus show the viability of our pursuit of modeling (ϵ, δ)-LDP and Differential Privacy.

3 Methodology

This work aims to explore the intersection of Information Theory (by the lens of Quantitative Information Flow), Privacy, and Fairness in Machine Learning. The primary goal is to investigate how these concepts can be effectively integrated and applied to ensure that machine learning models are transparent, robust, and fair, while maintaining the privacy of sensitive data.

First, we will conduct a comprehensive literature review to identify what is currently known and unkown about the relationship between the concepts of privacy, fairness and information flow in machine learning systems. The review will aim to highlight gaps in existing theoretical frameworks, particularly in terms of guarantees provided by common fairness and privacy metrics and notions.

Next, we will explore novel theoretical relationships among these areas. We are particularly interested in modeling the concept of $(\epsilon, \delta)-$Local Differential Privacy from the perspective of Quantitative Information Flow, as both the Differential Privacy and the Quantitative Information Flow frameworks propose methods of quantifying privacy, and there are already methods of modeling $\epsilon - LDP$ in the literature[10].

Additionally, we will explore the fundamental theoretical relationships between fairness and privacy, obtaining, for instance, the exact trade-offs between privacy and fairness constraints. The focus will be theoretical: by rigorous mathematical reasoning we aim to prove theorems that establish these relationships. We will also experiment with machine learning models trained on real-world datasets, testing empirically various privacy and fairness trade-offs, and identifying patterns that could lead to more robust theoretical results.

Finally, through the exploration of causality notions[19], we aim to provide new insights into the meaning of different quantitative notions of information flow, privacy, fairness, and of their relations, ultimately contributing to the creation of more transparent, accountable, and ethical AI systems.

4 Cronogram

1. 2025/1: Finish review of previous works. Coursework: Information Theory, Project and Analysis of Algorithms.

2. 2025/2: Model $(\epsilon, \delta) - LDP$ and DP with the QIF framework. Coursework: Deep Learning, Statistical Foundations of Data Science.

3. 2026/1: Explore theoretical relations between privacy and fairness metrics. Coursework: Measure Theory, Combinatorics.

4. 2026/2: Explore causality modeling for fairness and privacy.

Bibliography

[1] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 60–69. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/agarwal18a.html

[2] Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., Zhioua, S.: Survey on fairness notions and related tensions. EURO Journal on Decision Processes **11**, 100033 (2023). https://doi.org/https://doi.org/10.1016/j.ejdp.2023.100033, https://www.sciencedirect.com/science/article/pii/S2193943823000067

[3] Alvim, M.S., Andrés, M.E., Chatzikokolakis, K., Degano, P., Palamidessi, C.: On the information leakage of differentially-private mechanisms. Journal of Computer Security **23**(4), 427–469 (2015)

[4] Alvim, M., Fernandes, N., Nogueira, B., Palamidessi, C., Silva, T.: On the duality of privacy and fairness (extended abstract). In: International Conference on AI and the Digital Economy (CADE 2023). Institution of Engineering and Technology, United Kingdom (2023). https://doi.org/10.1049/icp.2023.2563, 9th International Conference on AI and the Digital Economy, CADE 2023 ; Conference date: 26-06-2023 Through 28-06-2023

[5] Alvim, M., Chatzikokolakis, K., McIver, A., Morgan, C., Palamidessi, C., Smith, G.: The Science of Quantitative Information Flow. Information Security and Cryptography, Springer International Publishing (2020), https://books.google.com.br/books?id=jJH-DwAAQBAJ

[6] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. URL https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing (2019)

[7] Cummings, R., Gupta, V., Kimpara, D., Morgenstern, J.: On the compatibility of privacy and fairness. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization. p. 309–315. UMAP'19 Adjunct, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3314183.3323847, https://doi.org/10.1145/3314183.3323847

[8] Dwork, C.: Differential privacy. In: International colloquium on automata, languages, and programming. pp. 1–12. Springer (2006)

[9] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. p. 214–226. ITCS '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2090236.2090255, https://doi.org/10.1145/2090236.2090255

[10] Fernandes, N., McIver, A., Sadeghi, P.: Explaining epsilon in local differential privacy through the lens of quantitative information flow. arXiv preprint arXiv:2210.12916 (2022)

[11] Fernandes, N., McIver, A., Sadeghi, P.: Explaining epsilon in local differential privacy through the lens of quantitative information flow. In: 2024 IEEE 37th Computer Security Foundations Symposium (CSF). pp. 419–432. IEEE (2024)

[12] Ferry, J., Aïvodji, U., Gambs, S., Huguet, M.J., Siala, M.: Sok: Taming the triangle - on the interplays between fairness, interpretability and privacy in machine learning. ArXiv **abs/2312.16191** (2023), https://api.semanticscholar.org/CorpusID:266573131

[13] Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. Commun.

ACM **64**(4), 136–143 (mar 2021). https://doi.org/10.1145/3433949, https://doi.org/10.1145/3433949

[14] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., Wallach, H.: Improving fairness in machine learning systems: What do industry practitioners need? In: Proceedings of the 2019 CHI conference on human factors in computing systems. pp. 1–16 (2019)

[15] Li, L., Lassiter, T., Oh, J., Lee, M.K.: Algorithmic hiring in practice: Recruiter and hr professional's perspectives on ai use in hiring. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 166–176 (2021)

[16] Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., Lin, Z.: When machine learning meets privacy: A survey and outlook. ACM Computing Surveys (CSUR) **54**(2), 1–36 (2021)

[17] Makhlouf, K., Zhioua, S., Palamidessi, C.: On the applicability of machine learning fairness notions. SIGKDD Explor. Newsl. **23**(1), 14–23 (may 2021). https://doi.org/10.1145/3468507.3468511, https://doi.org/10.1145/3468507.3468511

[18] Makhlouf, K., Zhioua, S., Palamidessi, C.: Survey on causal-based machine learning fairness notions (2022)

[19] Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, USA, 2nd edn. (2009)

[20] Pinzón, C., Palamidessi, C., Piantanida, P., Valencia, F.: On the incompatibility of accuracy and equal opportunity. Machine Learning (May 2023). https://doi.org/10.1007/s10994-023-06331-y, https://doi.org/10.1007/s10994-023-06331-y

[21] Roscher, R., Bohn, B., Duarte, M.F., Garcke, J.: Explainable machine learning for scientific insights and discoveries. Ieee Access **8**, 42200–42216 (2020)

[22] Santos, G., Figueiredo, E., Veloso, A., Viggiato, M., Ziviani, N.: Predicting software defects with explainable machine learning. In: Proceedings of the XIX Brazilian Symposium on Software Quality. pp. 1–10 (2020)

[23] Sheikh, M.A., Goel, A.K., Kumar, T.: An approach for prediction of loan approval using machine learning algorithm. In: 2020 international conference on electronics and sustainable communication systems (ICESC). pp. 490–494. IEEE (2020)

[24] Zhou, J., Joachims, T.: How to explain and justify almost any decision: Potential pitfalls for accountability in ai decision-making. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. p. 12–21. FAccT '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3593013.3593972, https://doi.org/10.1145/3593013.3593972