Relations between Fairness, Privacy and Quantitative Information Flow in Machine Learning

Type: Scientific
Advisor: Mário Sérgio Alvim

1st Artur Gaspar da Silva
Departamento de Ciencia da Computaçãp (DCC)
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil
artur.gaspar@dcc.ufmg.br

***Abstract*—Recent years have witnessed an enormous advance in the area of Machine Learning, reflected by the popularity of Artificial Inteligence systems. For most of the history of machine learning research, the main goal was the development of machine learning algorithms that led to more accurate models, but it is now very clear that there are many other important areas to develop. We want models to be fair to unprivileged groups in society, to not reveal private information used in the model training, to provide comprehensible explanations to humans in order to help identifying causal relationships, among many relevant goals other than simply improving model accuracy. In this work, we explore possible new relationships between fairness, privacy and Quantitative Inforamtion Flow. The first exploration is an analysis of papers that explore the impact of privacy-enhancing mechanisms on Machine Learning fairness notions. Our second exploration is the possibility of dividing a fixed local differencial privacy budget between variables with varying degrees of sensitivity. Finally, we explore modeling both local differential privacy parameters within the Quantitative Information Flow framework.**

***Index Terms*—Quantitative Information Flow, Differential Privacy, Fairness, Machine Learning.**

I. Introduction

Recent resarch [1] [2] [3] [4] 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 [5]. Also, there has been some work on introducing Causality concepts into the discussion, for instance, to develop better fairness metrics [6]. 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 [7]. 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 [8], loan approvals [9], hiring decisions [10], and others.

The goal of this Undergraduate Thesis is to review and reproduce results presented in the litearture, verify the viability of the connections between the aforementioned areas and Quantitative Information Flow, and, if possible, develop new theoretical results. This project is divided into two parts: POC I and POC II. In POC I, the specific goals were to research the literature for these concepts and focus on the connections that have been identified between them, so the expected result is a concise review of the literature on these topics. In this work (POC II), we verify possible connections between Privacy, Fairness and Quantitative Information Flow, and outline possible new theoretical results. We provide an in-depth theoretical analysis of the viability of Quantitative Information Flow approaches to these areas and the connections between privacy and fairness. We provide a formal exploration of the impact of privacy-enhancing obfuscation methods in fairness, based on important results in the literature reviewed in POC I, we explore how the privacy budget can be divided between many variables in the context of Local Differential Privacy, and, finally, we explore how viable is the application of the Quantitative Information Flow framework in Local Differential Privacy.

II. Theoretical Reference

Fairness in Machine Learning is concerned with measuring how unfair the results provided by Machine Learning models are to certain groups or individuals [11], and improving how fair the models are [12]. There are tensions between different fairness measures [13] [14]. 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 [15] 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]. 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 [17]. In recent research [1], the relationships between Fairness, Interpretability and Privacy have been extensively explored. Recent papers focuses on relationships between Privacy and Fairness [4], on the relationship between Privacy [3], and on the feasibility regions of Accuracy and Fairness metrics [5] [2].

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 [18]. Also, we have recent work [19] 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.

A. Machine Learning and scenario considered

Machine Learning is the field of study that focuses on developing methods of learning general patterns from limited data. In recent years, important advances have been observed in Machine Learning research, and also the popularity and applications of some methods have increased significantly. One example is the improvement of Convolutional Neural Network architectures, and the popularity of Generative Models. Also, some recent research is focused on the theory behind such machine learning methods [20] [21], and statistical learning in general [22]. Part of the goal of this formalization is to provide more qualitative guarantees in regard not only to accuracy, but also fairness, privacy, interpretability, and other important qualities. We discuss some of these different goals in the next two subsections.

In general, we consider *supervised learning* problems: in this scenario, a *machine learning algorithm* is an algorithm that receives many data points, which we call *training data*, and outputs a *model*. This model is itself an algorithm that receives a data point with some information omitted, encoded in what we call the *target variable*, and outputs a guess of the omitted information, which we call the *model prediction*. The model is then evaluated with other data points, ideally not the same ones used for training the model. This is called supervised learning because the algorithm has access to the target variable during the training process, which is not the case for unsupervised learning.

Figure 1 shows how the other concepts are related to machine learning in this context: we usually can assume that the training data is generated by some causal process, which can be modeled by a causal model; possible privacy attacks include performing sensitive information inference on the training data and on the model itself; we usually measure how accurate and fair a system is by analyzing its predictions for many data points; we can also obtain local and global explanations for complex models by this type of analysis. Throughout this wor, we focus on privacy and fairness, and how these concepts might be modeled within the Quantitative Information Flow framework.

B. Fairness

In the context of Machine Learning, fairness refers to the reduction, as much as possible, of *algorithmic bias*, the bias introduced by algorithmic decisions. This bias might have a big social impact because this can expand existing unfair discrimination in society, as machine learning algorithms are being used to make more and more important decisions. One famous example is the COMPAS recidivism algorithm, which has been used by the United States courts to estimate how likely someone is to reoffend in the future. It was revealed [8] that this tool was heavily biased against black people.

For binary classification, we will say that the result is *positive* for a data point if it benefits the person represented by that data point, and *negative* otherwise. We will say that the *unprivileged group* is the group of people affected negatively by the bias, and the *privileged group* is the other group of people.

Such biases can happen because of many factors. The algorithm itself might introduce bias, or the data might be biased. The data may have been collected in a biased way (in the COMPAS example, this would be the case if reincidivism data was collected more for black reincidivists than for white), or the data might be simply reinforcing some bias in society.

Also, the bias in society might be such that the data is in disagreement with reality (the unprivileged group's true values for the target variable would affect them in the same way as the privileged group), or it is in agreement with reality because of structural biases in society. For instance, if the prediction of the algorithm is whether someone will have good grades if accepted to some university, people in the unprivileged group might not have had as good oportunities in life as people in the privileged group, so the data is correct when it says that those people will have worse grades. Even so, the results might still be considered unfair, as this depends on the notion of fairness we consider. All of these unfairness scenarios can be further divided into other types of unfairness, as was done in previous works [23]. Image 2 illustrates where unfairness might come from, and we summarize below the ways in which unfairness might be introduced:

1) Algorithm results do not reflect the data.
    a) The algorithm might optimize for the majority only, achieving good overall accuracy even though it's mostly wrong for minorities. This can be considered a type of Aggregation Bias.
    b) Systematic errors in the algorithm, that lead to biased estimation.
2) The data can be biased, not reflecting the reality.
    a) We can have structural biases in society, such that people in unprivileged groups do not have the

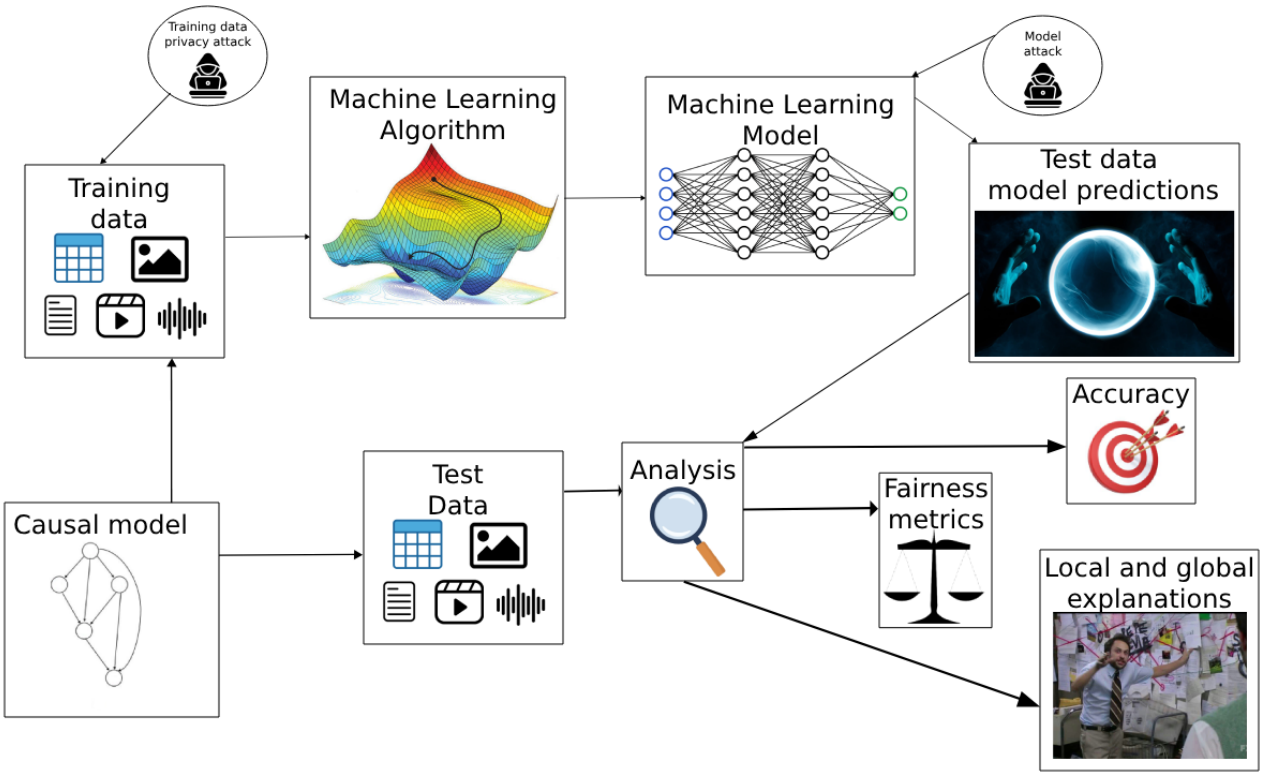

Fig. 1. Figure representing the supervised learning scenario we consider throughout this work, with a focus on privacy and fairness.

necessary opportunities, but if they were treated similarly to the privileged group by society, they would have similar results. For instance, an unprivileged group that does not have good educational opportunities will have worse scores on exams because of historical discrimination, and although just looking at whether someone is in this group could lead to a good accuracy, it might only perpetuate current unfair biases in society.

b) The bias can also be introduced in a way that people in the unprivileged group were misclassified before the data was collected. For instance, maybe capable people in an unprivileged group usually do not get a job even though they are actually as capable as the unprivileged group.

c) Data collection does not reflect the reality: Measurement bias (for instance, COMPASS used friend/family arrests as a proxy for a risk score present in the dataset), Omitted Variable Bias (this violates assumptions of some learning models, for instance linear regression models usually assumes error terms uncorrelated with the parameters considered in the regression), Representation/Sampling Bias (biased sampling lacking the diversity of the population), Simpson's Paradox (if we do not have data on a confounder, correlations might be spurious [24]).

d) If the data is collected on a group fundamentally distinct from the one where it will be used, for instance another population (Population Bias) or the same population but at another time (Temporal Bias), unfair bias might be introduced.

e) Data that relies on people's opinion is prone to many biases: Social Bias (people do what others are doing), Self-Selection Bias (people think that everyone agrees with them), and many others.

3) Data might depend on the algorithm's previous output: Presentation Bias (the user is presented to some selected advertisements, for instance), Ranking Bias (search engines ordering results in a biased way), Popularity Bias (more popular items are shown more). This might strengthen biases through time.

4) Finally, the circumstances can change through time, either by the influence of the algorithm itself or other factors, which can worsen the quality of algorithms previously considered able to provide good results (Emergent Bias).

Besides biased data and deliberate bias in the algorithm, such that the results of the algorithm do not reflect the data, it is also possible to introduce bias because the algorithm might prioritize making correct predictions for the majority of the population, if it can't make correct predictions for both the majority and the minority. Another possibility is that the prediction might depend on past decisions of the algorithm, and we only know the result if the result provided is positive (for instance, we only know if someone will reincide

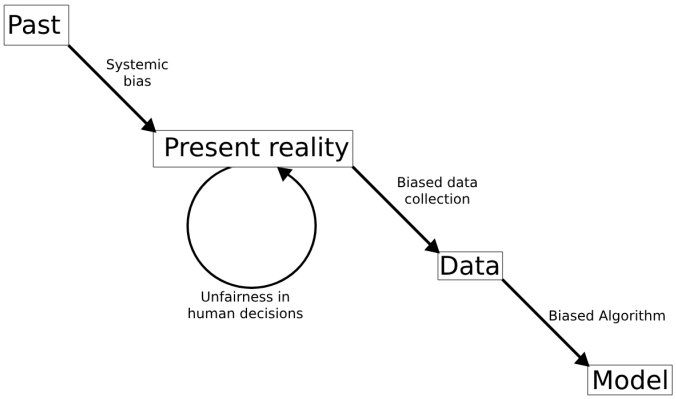

Fig. 2. Figure representing some of the main possible sources of unfairness.

if we release them). In this type of scenario, according to Learning Theory, it's important to take suboptimal decisions to *explore* different options and gather more data [25] (found in arxiv.org), which might be considered unethical as it might have a big cost to society (releasing someone that's probably going to commit more crimes) or to the individual (not giving a life-saving drug to some patients as an experiment to see the survival rates for that specific group).

Many different notions of algorithmic fairness have been developed, and some are not compatible [26]. Initially, the notions of fairness could be grouped into two main types: statistical and individual definitions of fairness [25]. Statistical (group) notions of fairness require some statistical metrics to be similar for certain demographic groups, and individual notions enforce constraints on pairs of individuals, for instance requiring similar individuals to be treated similarly. Many problems with statistical notions and why they, in general, do not provide good individual guarantees are presented in previous works [4] [27]. For instance, one such problem is satisfying the constraints for two protected attributes individually but not for combinations of these attributes. One problem with both individual and group notions is *composition*: it is not always the case that satisfying fairness constraints in individual, isolated, components of a system imply that fairness constraints will be satisfied for the whole system [28]. Finally, there are also causal approaches to fairness notions. In general, it is not possible to satisfy some of the main notions of fairness at the same time [29] [30] [31], and which fairness notion to use depends a lot on the specific goals of each different system.

We will now define some of the main notions of fairness, which are the ones we will consider in Subsection III-A. We consider that Y is the binary target variable, with $Y = 1$ as the positive result and $Y = 0$ as the negative one; A is the binary sensitive attribute, with $A = 1$ for the privileged group and $A = 0$ for the unprivileged group; $\hat{Y}$ is the model prediction of the target variable value; X is a set of legitimate factors that can be used for classification.

Definition 1 (Equal Opportunity Difference)**.** *We define* Equal Opportunity Difference *as* $P(\hat{Y} = 1|A = 1, Y = 1) - P(\hat{Y} = 1|A = 0, Y = 1)$. *Equal Opportunity is* satisfied *if the Equal Opportunity Difference is equal to zero.*

Definition 2 (Statistical Disparity)**.** *We define* Statistical Disparity*, also known as* Demographic Disparity*, as* $P(\hat{Y} = 1|A = 1) - P(\hat{Y} = 1|A = 0)$. *Statistical (Demographical)* Parity *is* satisfied *if the Statistical Disparity is equal to zero.*

Definition 3 (Conditional Statistical Disparity)**.** *We define* Conditional Statistical Disparity*, conditioned on* x*, as* $P(\hat{Y} = 1|A = 1, X = x) - P(\hat{Y} = 1|A = 0, X = x)$. *Conditional Statistical* Parity *is* satisfied *if the Conditional Statistical Disparity is equal to zero.*

A possible general definition of *individual fairness* notions is that an algorithm is considered fair if it gives similar outcomes to similar individuals, according to similarity notions relevant to the specific scenario considered.

The techniques developed to reduce unfairness in algorithmic decision-making can be divided into *pre-processing*, *in-processing* and *post-processing*. Pre-processing techniques modify the training data to remove biases present there. In-processing techniques modify the learning algorithm itself, for instance, by changing the objective learning function to include not only accuracy but also adding to it some statistical fairness metric, or including some constraint that it has to satisfy. Post-processing techniques act after the model is trained to reduce the unfairness in the decisions made by such a model.

In general, just removing the variables that would be considered unfair to use directly to classify an individual is not enough to guarantee fairness. As illustrated in image 3, the variables we would remove might be highly correlated to other variables, which could be used by the model to discriminate almost as if we hadn't removed any variable. Also, even if the machine learning model itself didn't use any sensitive variables or correlated attributes for the predictions, we still need to collect this sensitive data to be able to measure how unfair the model is.

C. Privacy

In the context of Machine Learning, a privacy-preserving algorithm is one that does not allow information considered private/sensitive to be obtained by unauthorized parties. According to the terminology presented in [16], this is called *private ML*. The private information to be protected can be the data used to train the model or the model parameters and structure itself. It is also possible to use Machine Learning to enhance privacy, *ML enhanced Privacy Protection*, or to serve as an attack tool, *ML-based Privacy Attack*. We will focus on private ML.

We call *adversary* the agent that wants to discover the private information, and *secret* the private information itself. There are some possible goals of the adversary: she might wish to recover the model itself by trying to approximate the function that represents the model, to recover some feature

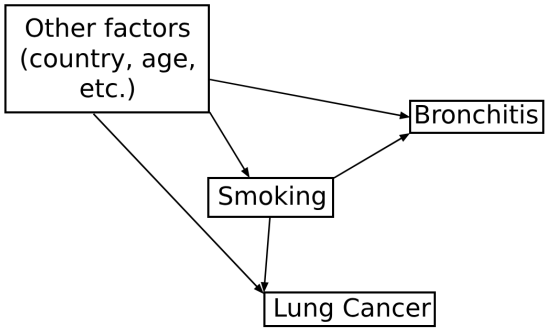

Fig. 3. Figure representing potential correlations between sensitive variables and other factors: if the disease status is a sensible information that could be used for unfair discrimination, then removing this information might not be enough to avoid unfair discrimination, as smoking, age and other factors combined might lead to unfair discrimination almost as if the model had direct access to the sensitive values.

or statistical property of the dataset, to discover whether some individual data point is present in the dataset, or even recover the exact values of individual samples in the dataset. We distinguish between the *White-Box access* and *Black-Box access* scenarios as the situation in which the adversary has or does not have full access to the trained model and their parameters, respectively.

Model Extraction Attacks assume an adversary with black-box access to the model, and no prior knowledge of the model parameters and training data. Some approaches to model extraction are presented in previous works [32]. Although the most efficient attacks rely on confidence values, attacks that rely only on the output class labels are also presented. Other works focused on estimating hyperparameters [33] for an adversary that knows the training dataset and the Machine Learning algorithm. Notice that recovering the model can help in the development of attacks against the training data, even if the adversary does not have prior knowledge of the model.

Feature estimation attacks focus on recovering statistical properties or features of the training dataset, for an adversary that does not know the training data or the its distribution. Some attacks are presented in previous works [34] both for black-box and white-box model access. As expected, the white-box attacks lead to better attacks, and they provide examples for recovering individual sensitive information from marital happiness answers in two different datasets and white-box attacks for recovering images in a face recognition model. The results lead to the possibility of (almost) identifying someone with only their name and the face recognition model and to the possibility of identifying the answer to a sensitive topic on a supposedly anonymous questionnaire (whether the person answering ever cheated on their significant other) with very high precision and a recall bigger than 20%.

The attacks presented in some previous works [35] aims to recover statistical information about the training dataset by the use of many models trained on different datasets and a meta-classifier that identifies if a given model was trained in a dataset with some desired statistical property. This is a white-box scenario, and they focused on attacking Support Vector Machines and Hidden Markov Models. Another common type of attack is the Membership Inference Attack, reviewed in previous works [36], in which the adversary aims to infer whether a given data point was used to train a given model or not. One approach is presented in a previous work [37], in which the adversary can sample from the original data distribution and has black-box access to the target model. It works by training many "shadow models" by using the data distribution, half with the target data point and half without it, then performing some computations based on confidence scores to estimate how likely the real model is to have used the relevant data point.

We will now focus on methods of modeling the protection of training data from privacy attacks. The main methods we will mention are Differential Privacy, Local Differential Privacy, and Homomorphic Encryption. Local Differential Privacy will be the method we focus on in Subsections III-A, III-C and III-C.

Differential Privacy (DP) is one of the most important definitions of quantifying privacy: the idea of DP is to define how hard it should be to distinguish one dataset from another that differs by at most any one individual data point. More generally, we can define how hard it should be to distinguish an element from a neighboring element, such that in the database example, the elements are databases and the neighboring relation is such that two databases are neighbors if and only if they differ by at most one data point. A (possibly randomized) algorithm that aims to obtain a dataset that is protected according to the DP definition is called a *Differential Privacy Mechanism*. The formal definition is presented below.

Definition 4 (ϵ-Differential Privacy)**.** *A randomized algorithm from $\mathcal{X}$ to $\mathcal{Z}$ satisfies ϵ-Differential Privacy, for $\epsilon > 0$, if for every $x, x' \in \mathcal{X}$ such that x is a neighgboring element of x', and for every $S \subseteq \mathcal{Z}$:*

$$P(\mathcal{K}(x) \in S) \leq e^{\epsilon} P(\mathcal{K}(x') \in S)$$

In some scenarios, we consider $S = \{y\}$, and the restriction of Definition 4 is equivalent to $P(\mathcal{K}(x) = y) \leq e^{\epsilon} P(\mathcal{K}(x') = y)$, which is equivalent to $P(Y = y|X = x) \leq e^{\epsilon} P(Y = y|X = x')$ if we represent the output of $\mathcal{K}$ as the random variable Y. This alternative definition can be used for other definitions of Differential Privacy and even Local Differential Privacy, and in the last subsection of the contributions part of this work we will use this alternative definition. This choice was made because it simplifies the results and proofs, and the attempts of the referred contributions subsection were ultimately unsuccessfull, so there is no reason to expand it to the most general version possible if it already does not work in the simpler version.

The parameter ϵ can be interpreted as how close we require the probabilities of neighboring datasets to be, such that smaller values of ϵ lead to stronger requirements. In some practical applications, the Differential Privacy restriction is

too strong. Thus, we have a relaxed definition for Differential Privacy, in which the parameter δ can be interpreted as the probability that the DP guarantee will not be satisfied:

Definition 5 ((ϵ, δ)-Differential Privacy)**.** *A randomized algorithm from $\mathcal{X}$ to $\mathcal{Z}$ satisfies (ϵ, δ)-Differential Privacy, for $\epsilon > 0$ and $\delta \in [0, 1]$, if for every $x, x' \in \mathcal{X}$ such that x is a neighgboring element of x'.*

$$P(\mathcal{K}(x) \in S) \leq e^{\epsilon} P(\mathcal{K}(x') \in S) + \delta$$

Local Differential Privacy (LDP) is a concept important when the data collector is not assumed to be trusted and consists of the same restrictions as Differential Privacy, but with $\mathcal{X}$ as the set of possible individual values, $\mathcal{Z} = \mathcal{X}$ and the neighboring relation being such that all possible values are neighbors. This algorithm should be applied locally by the owner of each data point.

The idea is that individuals apply noise locally in a way such that the data collector can still obtain the desired results with the noisy data. The classical example is the scenario in which we want to discover how many people do some illegal activity (for instance, use illegal drugs) in a given region. The person answering might be tempted to lie to not let this sensitive information leak. But if we tell them to toss two coins, such that if the first one comes up heads, they answer truthfully, but if not, then the answer should be "yes" if the second coin came up heads and "no" otherwise. We will know that approximately $\frac{1}{2}$ of the answers is not the true answer of the person, such that half of these are "yes" and half "no". We can then remove 25% of the total number of answers from the number of "yes" and another 25% from the total number of "no" to estimate the real distribution. Previous works [38] delves into many mechanisms, information metrics, and applications relevant to Local Differential Privacy, including machine learning on private data.

There are some common misconceptions about what exactly are the assumptions of Local Differential Privacy. Notice, for instance, that if the data points of two individuals are known to be highly correlated (for instance, genetic data for two siblings), then even if their data points after the application of the LDP mechanism do satisfy (δ, ϵ)-LDP, the tuple of these two data points may not satisfy (δ, ϵ)-LDP, which can improve the inferences that an adversary can make. Imagine the extreme case: if we know that all data points are equal and the LDP mechanism gives a higher probability of not changing the data point, then the adversary can discover the common value of all data points simply by looking at the most common value after the mechanism is applied. This can also be a problem for Differential Privacy, as shown in previous works [39].

The possible dependencies among data points have led to some different definitions of what exactly are the assumptions of LDP, for instance, that all data points are independent, or that the adversary knows all data points but one. Previous works [40] discuss how a causal interpretation can help in uncovering the meaning of each LDP assumption, which are or not equivalent, and also compares with potential causal notions of LDP.

Federated Learning is another method that can help improve the privacy of individuals. The idea is that each individual trains a Machine Learning model locally, and shares information with a centralized server to improve a global model. The privacy risks are reduced because no individual data point and no individual user updates to the model are stored in the server. But still, without extra preparations, it might be possible to attack individual data points, as explored in previous works [41].

Finally, *Homomorphic Encryption* is a form of encryption that allows computations to be done without decrypting the data, just the result is decrypted. *Secure Multi-Party Computation* is also an option if there are multiple parties responsible for this computation. The major drawback of these methods is the significant additional computational cost. Homomorphic Encryption and Secure Multi-Party Computation have been proposed for frequency estimation [42], Deep Learning [43] [44], and others.

D. Quantitative Information Flow

The area of Quantitative Information Flow deals with methods of quantifying information leakage from systems. This estimation is important to consider when developing real systems, as some information leakages are acceptable. For instance, whenever someone tries to authenticate with a username and password, but incorrectly guesses the password, some information leaks about the real value of the password: we now at least know that it's not the one that they tried. But we usually agree that this is acceptable while revealing the real password whenever someone makes an incorrect guess is unacceptable. How to adequately quantify the amount of information leaked from a system might depend on the goals of the people involved and on the information they have before the system executes. We thus need to first define some important notions, illustrated by Figure 4, before proceeding:

1) *Adversary* is an agent that tries to gain something with the information that leaks from the system.
2) *Secret* is the non-public data that the system processes.
3) *Prior Disribution* is the probability distribution on secrets that represents the knowledge of the adversary before the system runs, how likely the secret is each possible value according to the adversary's prior knowledge.
4) *Posterior Distribution* is the hyper-distribution, a probability distribution on probability distributions, on secrets. It represents the knowledge of the adversary after the system runs: ignoring some technicalities, we can consider this hyper-distribution to have one distribution per observable system output representing the adversary's knowledge of the secret for each possible observable system output.
5) A information-theoretical *Channel* is a representation of the system, which encodes the distributions of possible

observable values output from the system, which might depend on the secret value.

6) The set $\mathcal{X}$ represents the set of possible values of the secrets.
7) The set $\mathcal{Y}$ represents the possible values of observable outputs of the system.
8) The set $\mathcal{W}$ represents the possible values of actions the adversary might take.

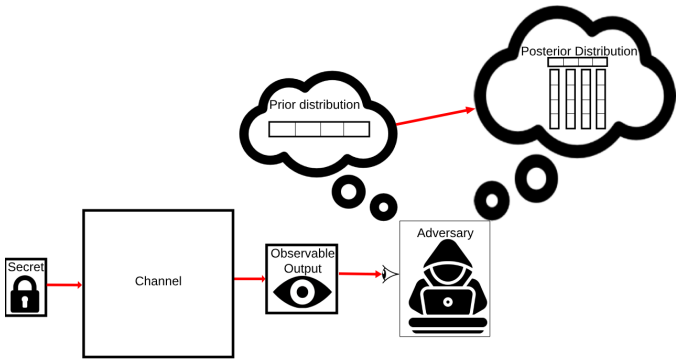

Fig. 4. Figure representing the scenario we consider in Quantitative Information Flow.

We consider the g-vulnerability framework, introduced in (at the time of writing this work) main QIF textbook [17]. This framework introduces new definitions, which we present more formally:

Definition 6 (Gain Function)**.** *A* Gain Function *is a function* $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$*, such that* $g(w, x)$ *defines the gain of the adversary if she takes the action* w *when the secret value is actually* x*.*

We do not have a loss for the system owner because we consider a zero-sum game: the gain of the adversary is exactly the loss of the people responsible for the system.

Definition 7 (Prior g-Vulnerability)**.** *The* Prior Vulnerability *of the system is defined as the average gain of the adversary if she takes the action that maximizes her expected gain, according to the distribution on secrets that represents her prior knowledge. Given a prior distribution on secrets* π *and a gain function* g*, this is defined as:*

$$V_g(\pi) = \max_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} \pi_x g(w, x)$$

Definition 8 (Posterior g-Vulnerability)**.** *The* Posterior Vulnerability *of the system is defined in the same way as the prior vulnerability, but considering the hyper-distribution that represents the posterior knowledge. Given a prior distribution on secrets* π *and a channel* C *and a gain function* g*, this is defined as:*

$$V_g[\pi \triangleright C] = \sum_{y \in \mathcal{Y}} \max_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} \pi_x C_{x,y} g(w, x)$$

The posterior g-vulnerability represents what will be the expected gain of the adversary after the system runs, but the estimative is made according to the knowledge the adversary has before the system runs

Definition 9 (Additive Leakage)**.** *The* Additive Leakage *can be defined as the difference between posterior and prior vulnerability:*

$$\mathcal{L}_g^+(\pi, C) = V_g[\pi \triangleright C] - V_g(\pi)$$

Definition 10 (Multiplicative Leakage)**.** *The* Multiplicative Leakage *is the result of dividing the posterior vulnerability by the prior vulnerability values:*

$$\mathcal{L}_g^\times(\pi, C) = \frac{V_g[\pi \triangleright C]}{V_g(\pi)}$$

There are many valuable theoretical results about channels regarding the relationships between prior and posterior g-vulnerabilities. We mention some of these results:

1) The chapter 7 of the main QIF textbook at the time of this writing [17, Chapter 7] shows results about the *capacity* of a channel, which is the maximum possible (additive or multiplicative) leakage that can happen through a channel if we fix either the prior, the gain function or neither.
2) Chapter 9 of the same book [17, Chapter 9] shows results about *refinement* of channels: in short, a channel is strictly better (for all priors and gain functions) than another channel in respect to the posterior vulnerability if and only if it can be written as a post-processing of this other channel.
3) Chapter 10 of the same book [17, Chapter 10] presents the notion of *Dalenious vulnerability*: it might be the case that the adversary is interested in a secret other than the one considered in the system, and can obtain information about this other system via a known joint distribution between this other secret and the secret that the system considers. In this case, a channel is also strictly better than another in respect to Dalenious leakage, for any such joint distribution and gain function, if and only it can be written as a post-processing of this other channel.
4) Finally, chapter 11 [17, Chapter 11] discusses the axiomatic characterization of the notion of vulnerability, and even how some results can be obtained by different axioms that consider the worst-case scenario instead of the average gains of the adversary. These axioms are explained in Definition 11 and 12, respectively.

Definition 11 (MAX QIF axiom)**.** *Given a Hyper distribution on secrets (a distribution on a distribution of secrets)* $\triangle \in \mathbb{D}^2(\mathcal{X})$*, the posterior vulnerability can be computed as the supremum of the vulnerability of all inner distributions of the hyper that have more than* 0 *probability. If there is a finite number of inner distributions of secrets in the hyper, this can be written as:*

$$V(\triangle) = \max_{\pi \in \triangle : \triangle_\pi > 0} V(\pi)$$

Such that $V(\pi)$ is the vulnerability value assigned to $\pi \in \mathbb{D}(\mathcal{X})$.

Definition 12 (AVG QIF axiom)**.** *Given a Hyper distribution on secrets (a distribution on a distribution of secrets) $\triangle \in \mathbb{D}^2(\mathcal{X})$, the posterior vulnerability can be computed as the expected vulnerability of all inner distributions of the hyper. If there is a finite number of inner distributions of secrets in the hyper, this can be written as:*

$$V(\triangle) = \sum_{\pi \in \triangle} \triangle_\pi V(\pi)$$

Such that $V(\pi)$ is the vulnerability value assigned to $\pi \in \mathbb{D}(\mathcal{X})$.

Even though most of the work on Quantitative Information Flow was developed with a stronger focus on measuring how system a system is, the g-vulnerability framework can be considered a general notion for quantifying information flow. This means that, in the future, we might be able to use it in the areas explored in this work, as some can be viewed in terms of information flows.

E. Fairness × Privacy

Some impossibility results in the literature argue why it is impossible to achieve both privacy and group fairness constraints under non-trivial accuracy [3]. We also have positive results that show how stasifying privacy constraints can help mitigate group unfairness [45] [46] [47], and we also have similar positive results results for individual fairness notions [4].

1) "A systematic and formal study of the impact of local differential privacy on fairness: Preliminary results" [45] aims to provide theoretical results that relate how training a machine learning model on a dataset in which local differential privacy mechanisms were applied can impact the fairness of this model if the data-gatherer does not try to reverse the applied noise. The results are provided for a simplified machine learning model, from a theoretical perspective.
2) "On the impact of multi-dimensional local differential privacy on fairness" [46] aims to evaluate experimentally how training a machine learning model on a multi-dimensional data set in which local differential privacy mechanisms were applied can impact the fairness of the model, again if the data-gatherer does not try to reverse the noise.
3) "(local) differential privacy has no disparate impact on fairness." [47] executes an empirical evaluation of the impact of many Local Differential Privacy mechanisms on fairness when we do not try to reverse the applied noise, including a new privacy budget allocation scheme based on the domain size of sensitive attributes.
4) "Fairness through awareness" [4] introduces a notion of individual fairness that is a generalization of differential privacy and explores the relationship between this notion of fairness, differential privacy, and statistical disparity.
5) "On the compatibility of privacy and fairness" [3] presents theoretical results that show the impossibility of having a classifier that is not trivially accurate and at the same time satisfies $(\epsilon, 0)$-differential privacy and equality of opportunity constraints. The paper also shows a PAC learner for an approximate fairness definition they provide.
6) "Exploring fairness and privacy in machine learning" [48] compiles results that relate the concepts of privacy and fairness, and also causal discovery.
7) "Toward learning trustworthily from data combining privacy, fairness, and explainability: an application to face recognition" [49] discusses the use of Homomorphic Encryption in combination with fair representation learning [31] to provide both privacy and fairness guarantees, respectively, when training machine learning models. Then the paper discusses how to provide local and global explanations for the model.

F. Fairness × QIF

One possibility of measuring fairness with Quantitative Information Flow notions is to measure the *reverse flow*: instead of looking at how much observing the output of a model helps in estimating input values (this would be a privacy concern), we measure how much observing the sensitive attribute helps in estimating output values. This idea is already explored in previous works [50] [51], so we won't explore it in this work.

G. Privacy × QIF

As QIF aims to quantify the flow of information, we can naturally consider applying it to the privacy scenario, by measuring how much information leaks about an arbitrary individual. One general idea is that differential privacy is a worst-case notion and g-vulnerability is an average-case notion, and there are results [17] showing that differential privacy implies bounds on leakage under arbitrary gain functions and prior distributions, but not the opposite.

1) "On the information leakage of differentially-private mechanisms" [18] discusses 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.
2) "Explaining epsilon in local differential privacy through the lens of quantitative information flow" [19] discusses 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.

III. Contributions

Besides the literature review, this work has three main theoretical contributions, which we cover in the three subsections of this section:

1) We identify qualitative questions about the premises of recent papers on the impact of Local Differential Privacy obfuscation mechanisms on fairness metrics, and thus questioning the conclusions obtained.
2) We analize the possibility of distributing noise among variables with different degrees of sensibility for the reduction of the total noise in Local Differential Privacy obfuscation mechanisms, and conclude that this is not a good idea for various reasons we discuss in the second subsection of this section.
3) We try a new approach, inspired on a recent paper [19] that models the ϵ parameter of ϵ-LDP, to modeling the δ parameter of (ϵ, δ)-LDP within the QIF framework, and conclude that it is not possible to do with this approach.

A. The impact of Local Differential Privacy mechanisms on fairness metrics

Recent theoretical and empirical investigations suggest that the application of Local Differential Privacy (LDP) obfuscation mechanisms may intrinsically mitigate the unfairness of machine learning models trained on noisy datasets [45] [46] [47]. This subsection proceeds as follows: First, we present a detailed analysis of relevant results from key recent studies. Subsequently, we discuss strong methodological limitations observed in these studies.

One of the recent studies on this topic [45] focuses on theoretical results. The assumptions and simplifications made are:

1) There is only one sensitive and one non-sensitive input variables, and one output binary variable that the model tries to predict, all binary.
2) The LDP obfuscation mechanism is Randomized Response, but it is applied only on the sensitive variable.
3) The fairness metrics considered are Statistical Disparity (SD) 2, Conditional Statistical Disparity (CSD) 3 and Equal Opportunity Difference (EOD) 1.
4) There is enough training and testing data to adequately estimate the real data distribution.
5) When given an input (a, x), the model outputs the most common value of Y seen in the training data distribution conditioned on $A = a, X = x$.
6) The model is trained directly on the noisy data distribution, we assume that the data collector will not try to reverse the noise and approximate the real data distribution.
7) The model is tested in the real data distribution, without noise.

The theoretical results obtained show that Conditional Statistical Disparity 3 and Equal Oportunity Difference 1 values remain equal or are reduced wihout reversing the signal. Also, Statistical Disparity 2 remains equal or is reduced, but the signal might be reversed. This means that Statistical Disparity never gets worse for the unpriviledged group, but might end up even stronger than before, but in favor of the unpriviledged group.

Another recent paper [46] analyzes the same scenario, but with more than one sensitive variable and different obfuscation algorithms (multi-dimensional LDP). The paper considers Randomized Response applied individually to each variable, and the version of Randomized Response from Definition 13, but considering all variables as one variable in a tuple format.

The results are empirically, not theoretically, verified, and they reach the same conclusions as the previously mentioned paper [45]. One extra analysis is how the distribution of outputs for each group can affect how much each group is affected by the obfuscation, so they are usually not equally impacted by the obfuscation mechanisms verified.

The last study we will consider in this work [47] also verifies the impact of obfuscation mechanisms on the fairness of models trained in the noisy dataset. The main difference is that they consider an approach that applies different ammounts of noise according to how many possible values each variable has. Also, this study focuses on extensive empirical, instead of theoretical, evaluation.

The results include which methods provided the best utility-fairness trade-off. The paper also discusses how their proposed method of splitting the privacy budget according to how many possible values each variable has.

Now, we will discuss the main limitations with the approaches of these studies.

First of all, applying noise to a single variable poses privacy risks. Unless we have complete trust in the data collector (in which case, applying noise is unnecessary), they could potentially reverse the noise and leverage a model trained on the unperturbed distribution to infer the sensitive variable's value from the remaining data. The second part of this contributions section details the issues arising from applying noise solely to the sensitive variable and distributing noise between its correlated counterparts.

Even when noise is applied equally to all variables, it is still not reasonable to assume that the data collector will not reverse this noise and later train a biased model. This action, differently from applying nosie on a single variable, would not undermine the very purpose of noise application, which is to protect sensitive information, as LDP guarantees would hold. But it would still remove completely the fairness guarantees, as they depend on trusting the data collectors. So, this approach would require trust in them for fairness purposes even if we do not trust them for privacy purposes.

Then, if the data collector is considered untrustworthy in general, applying noise might become useless, as they can simply undo the perturbation (given enough data) and proceed with possible unfair intent. Thus, unless we trust in the data collector being fair even though we do not trust them for respecting user privacy, the results presented in the analyzed studies do not hold, and alternative methods that do not rely on the collector's integrity should be considered.

Notice that trust on the data collector is not simply trust that the people and company are not malicious, it also includes trusting that the data they pocess will not leak to malicious parties. This might happen through cibersecurity breaches, for instance, in which the data could be used to make unfair "data-based" decisions.

Finally, even under the assumption of complete trust in the data collector not attempting to reverse the perturbation when training the model and with noise applied to all variables for privacy preservation, the act of not reversing the noise simply constrains the trade-off options available to the collector between fairness and accuracy. By introducing noise, a fundamental tension is created: increasing noise enhances privacy but typically reduces the accuracy of any model trained on the perturbed data [52]. This tension is also present between fairness and accuracy, especially in scenarios in which the unpriviledged group is significantly evaluated worse in the training data [5]. If we use noise as a pre-processing method for improving fairness, then we are not considering other, possibly better options, that lead to better fairness and accuracy [2] [53].

B. How to distribute the Local Differential Privacy budget between variables with varying levels of importance

One simple mechanism to satisfy the Local Differential Privacy constraints of Definition 4 is *Randomized Response*, described in 13.

Definition 13 (n-Randomized Response)**.** *Let $|\mathcal{X}| = n$. Then, the Randomized Response Mechanism $\mathcal{K}$ that maps from $\mathcal{X}$ to $\mathcal{X}$ can be defined as:*

$$\mathcal{K}(x) = \begin{cases} x & \text{with probability } \frac{e^\epsilon}{n-1+e^\epsilon}, \\ y & \text{with probability } \frac{1}{n-1+e^\epsilon}. \end{cases}$$

For any values $x \neq y$ such that $x, y \in \mathcal{X}$.

Notice that this adds up to one: $\frac{e^\epsilon}{n-1+e^\epsilon} + (n-1)\frac{1}{n-1+e^\epsilon} = 1$. Also, this mechanism satisifies the $\epsilon - LDP$ restriction from Definition 4, as we show in Lemma 1.

Lemma 1. *n-Randomized Response (Definition 13) satisifes $\epsilon - LDP$ (Definition 4).*

Proof. Let S be any subset of the set of all outputs of $\mathcal{K}$. Notice that, by the definition of $\mathcal{K}$, this is a subset of secrets. Also, let $x \in \mathcal{X}$ and $\epsilon > 0$. If $x \in S$, then $P(\mathcal{K}(x) \in S) = \frac{|S|-1+e^\epsilon}{n-1+e^\epsilon}$, otherwise $P(\mathcal{K}(x) \in S) = \frac{|S|}{n-1+e^\epsilon}$. Now let $x, x' \in \mathcal{X}$ be any pair of secrets. We will divide the proof into five parts:

1) If $S = \emptyset$, then $P(\mathcal{K}(x) \in S) = P(\mathcal{K}(x') \in S) = 0$, so trivially $P(\mathcal{K}(x) \in S) \leq e^\epsilon P(\mathcal{K}(x'))$.
2) If $x \in S$, $x' \in S$, then $P(\mathcal{K}(x) \in S) = \frac{|S|-1+e^\epsilon}{n-1+e^\epsilon} \leq e^\epsilon \frac{|S|-1+e^\epsilon}{n-1+e^\epsilon} = e^\epsilon P(\mathcal{K}(x') \in S)$, as $e^\epsilon > 1$ when $\epsilon > 0$.
3) If $x \in S$, $x' \notin S$, then $P(\mathcal{K}(x) \in S) = \frac{|S|-1+e^\epsilon}{n-1+e^\epsilon} = e^\epsilon \frac{\frac{|S|-1}{e^\epsilon}+1}{n-1+e^\epsilon} \leq e^\epsilon \frac{|S|-1+1}{n-1+e^\epsilon} = e^\epsilon \frac{|S|}{n-1+e^\epsilon} = e^\epsilon P(\mathcal{K}(x') \notin S)$, as $e^\epsilon > 1$ when $\epsilon > 0$.
4) If $x \notin S$, $x' \in S$, then $P(\mathcal{K}(x) \notin S) = \frac{|S|}{n-1+e^\epsilon} \leq \frac{|S|-1+e^\epsilon}{n-1+e^\epsilon} \leq e^\epsilon \frac{|S|-1+e^\epsilon}{n-1+e^\epsilon} = e^\epsilon P(\mathcal{K}(x') \in S)$.
5) If $x \notin S$, $x' \notin S$, then $P(\mathcal{K}(x) \notin S) = \frac{|S|}{n-1+e^\epsilon} \leq e^\epsilon \frac{|S|}{n-1+e^\epsilon} = e^\epsilon P(\mathcal{K}(x') \notin S)$, as $e^\epsilon > 1$ when $\epsilon > 0$.

□

Notice that all probabilities are inversely proportional to n, but n increases exponentially in respect to the number of variables for each data point. This means that in real applications with many attributes, the probabilities might get prohibitively small. Inevitably this leads to exponentially more data being necessary to reach statistically significant conclusions about it when we increase the number of variables collected per data point. One possible improvement in respect to this problem is obfuscating only the privacy-sensitive variables. For instance, in the scenario described at Figure 5, it might be enough from a privacy perspective to obfuscate only 2 out of 6 the variables presented.

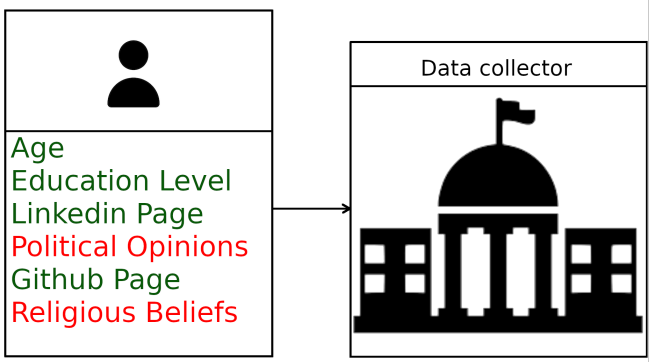

Fig. 5. Scenario in which some institution collects data on some sensible and non-sensible variables.

Applying obfuscation mechanisms at only a few variables might weaken the real privacy guarantees obtained. In the example illustrated by figure 5, for instance, the Age variable might be correlated with the Religious Beliefs variables. In this situation, applying noise just to the Religious Beliefs variable is not equivalent to applying the same noise in the scenario in which only the Religious Beliefs variable is available. In other words, even though an obfuscation mechanism provides guarantees in general, these guarantees are not kept if we apply noise only on one variable. For instance, Randomized Response (Definition 13) might guarantee ϵ-LDP if the noise is applied to all variables, as shown in Lemma 1, but not if it is applied to a subset of variables.

In the most extreme case, we have the toy example in which two variables A and X, such that A is considered sensible and X is not, and A is proportional to X. Then if we apply noise to A only, the data collector can discover the exact value of A of any individual by looking at the value of X for that individual. More realistic examples might have a sensitive variable A and a tuple of non-sensitive variables $(X_1, X_2, ..., X_n)$, such that the value of the tuple is highly correlated with A, or can be used to predict A with more than acceptable accuracy.

One important privacy paradigm to consider is that a system can be considered non-private if it leaks more information about the sensitive data than an adversary had *before* observing the system. Thus, one might think that the data collector would be unable to identify the strong corrleations mentioned if enough noise is applied. But we must keep in mind that many Local Differential Privacy mechanisms are designed such that the data collector can reverse the noise with some confidence if enough data is provided. For instance, in the Randomized Response Mechanism from Definition 13, we can infer that a fraction of the individuals will provide incorrect information about their data, and by some simple computations we can conclude that to (approximately) reverse the noise we can simply subtract a fraction $\frac{1}{n-1+e^\epsilon}$ of each obtained proportion of individuals with that value. Then, we re-scale the values obtained to add up to one.

Many LDP mechanisms allow for some method for the data collector to approximately reversing the noise, which allows to estimate the real data distribution, so the data collector can, in general and with enough data, approximate the real correlations between variables. This means that even if we consider the paradigm of comparing the prior and posterior information, applying noise to a few of the variables is still a big problem. A data collector with no prior information can discover that a set of non-sensitive variables is highly correlated with a sensitive variable, and use this to infer the value of the sensitive variable of any individual.

Even though it is not safe to apply noise to the sensitive variables only, we might still wish to find another strategies that protect the sensitive variables but not necessarily the non-sensitive variables. One way to do this is to apply more noise to the sensitive variables, and less noise to correlated variables, such variables with less correlation have less noise applied to them. In order to be able to quantify this idea, we provide Definition 14.

Definition 14 (Variable-ϵ-Local-Differential Privacy)**.** *A randomized algorithm from $\mathcal{X}$ to $\mathcal{X}$, such that $A \times B = X$ with A as the set of sensible variables and B as the set of non-sensible variables, satisfies Variable-ϵ-Differential Privacy, for $\epsilon > 0$, if for every $a, a', c \in A$ and $b, d \in B$:*

$$P(\mathcal{K}((a,b)) = (c,d)) \leq e^\epsilon P(\mathcal{K}((a',b)) = (c,d))$$

Definition 14 implies that an algorithm satisfies Variable-ϵ-Differential Privacy if, given any output, there is an e^ϵ degree of uncertainty about the value of the sensible variable before the application ofthe algorithm to the input data point. Also, note that A, B can themselves be defined as tuples of variables, and thus we can model any scenario with sensible and non-sensible variables this way.

The main reason why we might not be able to guarantee that the restriction from Definition 14 is satisfied is that the data distribution is usually not available before the data from many individuals is collected and aggregated. This is the most common scenario, as if the distribution of the variables is already known, then there is no reason at all to collect data for most applications. Then, it might not be possible to apply noise proportionally to how correlated a variable is to a sensitive variable, as the person applying the noise does not know all the correlations a priori. If we want the mechanism to work in any distribution, we need to apply the noise to all variables, as it must work even if other variables are equal to the sensitive variables.

Finally, one common idea is to simply not collect the sensitive data. The problem with this approach is that without this information, it is not posslbe to infer how unfair a system is. Verifying biases and general unfairness is the goal of many important studies, and thus collecting sensitive data, at least in an aggregated way that does not allow the identification of the sensitive information of specific individuals, is fundamental. The sensitive data collected by companies can, for instance, be used by regulatory agencies to identify unfair biases that the copany presents towards unpriviledged groups of customers and employees, for instance.

C. Modeling the δ parameter of (ϵ, δ)-LDP within the Quantitative Information Flow framework

Recent research [54] shows how to represent and explain the ϵ parameter of ϵ-LDP 4 through the lens of Quantitative Information Flow. They consider the alternative MAX axiom of QIF [17, Chapter 11], and the main result is to prove that the supremum of the multiplicative leakage of a given channel is limited by e^ϵ if and only if the fraction of the values of any two entries of the channel in the same column are also limited by e^ϵ. This can be, in turn, shown to be equivalent to the restriction of the definition ϵ-LDP 4.

Intuitively, one possible observation is that we are considering the worst-case scenarios when we use the MAX 11 instead of the AVG 12 axiom. Previous attempts were made to relate ϵ-LDP directly with QIF, but they relied on the more common AVG axiom for QIF [18], so a bound on LDP in terms of multiplicative leakage were not available, only bounds on leakage in terms of LDP parameters.

The idea of this subsection then, was to explore the usage of another alternative axiom, *δ-MAX* 15, instead of MAX 11 or AVG 12. We do not prove that δ-MAX has the same desired properties as AVG and MAX, we just applied it to the scenario considered and tried to reach significant conclusions. We will describe why one might think this idea could work, and outline the most promising line of trial and error we had, even though we utlimately reached negative results.

Definition 15 (δ-MAX QIF axiom)**.** *Given a Hyper distribution on secrets (a distribution on a distribution of secrets) $\triangle \in \mathbb{D}^2(\mathcal{X})$, the posterior vulnerability can be computed as the supremum of the vulnerability of all inner distributions of the hyper that have more than δ probability. If there is a finite number of inner distributions of secrets in the hyper, this can be written as:*

$$V(\triangle) = \max_{\pi \in \triangle : \triangle_\pi > \delta} V(\pi)$$

Such that $V(\pi)$ is the vulnerability value assigned to $\pi \in \mathbb{D}(\mathcal{X})$.

The key distinction between ϵ-LDP 4 and (ϵ, δ)-LDP 5 lies in how these privacy measures address worst-case scenarios. ϵ-LDP tolerates a zero probability of failure, while (ϵ, δ)-LDP allows for a δ probability of failure. Similarly, QIF under the MAX axiom 11 captures worst-case leakage with a non-zero probability, and QIF under the MAX-δ axiom 15 captures the worst-case leakage with a probability greater than δ.

This parallel suggests a potential relationship between these measures. Specifically, it may be feasible to establish a bound on (ϵ, δ)-LDP in terms of the parameters of QIF concepts under the δ-MAX axiom, mirroring the existing modeling of ϵ-LDP in terms of the multiplicative leakage of QIF under the MAX axiom. This approach could provide a more generalized connection between the QIF and the Differential Privacy frameworks. This extension is particularly relevant in settings where a small probability of privacy breach is acceptable, which is, in fact, a common setting.

Many attempts were made, but we will focus on the main one. First, we simplified the main theorem of the mentioned paper [54]. Then we wrote the version that should be true if our idea of using the MAX-δ axiom could lead to the desired results by using the same ideas from the demonstrations in the mentioned paper, and proved that it can't be proved the same way.

First, we will present the theorem as it was presented in the analyzed paper [54] (Theorem 1), and a simplified version that is more suited to our needs (Theorem 2).

Notice that a review of these results might be important if this proves relevant in the future, as we consider concrete instead of abstract versions of the channels throughout our arguments. If only abstract channels are considered, maybe our counter-example for this line of reasoning does not hold anymore.

Theorem 1 (Alternative bound to ϵ-LDP)**.** *Let $C : \mathcal{X} \to \mathbb{D}(Y)$ be any channel and $\epsilon \geq 0$. Also, let $\mathcal{X}$ be the set of secrets, Y be the set of channel outputs and J be the distribution of $\mathcal{X} \times Y$. Then:*

$$\sup_{\substack{\pi \in \mathbb{D}(\mathcal{X}): \\ \pi_x > 0}} \max_{\substack{x \in \mathcal{X}, y \in Y: \\ J_{x,y} > 0}} \frac{C_{x,y}}{p(y)} \leq e^{\epsilon}$$

If and only if, for any $x, x' \in \mathcal{X}, y \in Y$ such that $J_{x',y} > 0$ and $J_{x,y} > 0$ we have:

$$\frac{C_{x,y}}{C_{x',y}} \leq e^{\epsilon}$$

The first expression of Theorem 1 is equivalent to QIF notions of multiplicative capacity under the MAX axiom 11, so it effectively establishes a clear connection between the QIF and LDP frameworks. This equivalence and the proof of Theorem 1 are already presented at the paper [54], so we will not discuss it here. Also, we added the requirement of $J_{x',y} > 0$ and $J_{x,y} > 0$ to the theorem so there are no divisions by zero (as we can change x and x' and the theorem should hold, this is the same as requiring $J_{x',y} > 0$ only). Finally, we already expanded the terms defined in the paper [54], so we do not need to re-define everything again.

One important point to notice is that $C_{x,y} = P(Y = y|X = x)$, and thus the expression $\frac{C_{x,y}}{C_{x',y}} \leq e^{\epsilon}$ is equivalent to the restriction of the definition of ϵ-LDP 4 when $|S| = 1$.

Theorem 2 (Alternative version of ϵ-LDP bounds)**.** *The following logical affirmation is equivalent to Theorem 1, when channels are not allowed to have entries with value zero.*

For any channel $C : \mathcal{X} \to \mathbb{D}(Y)$ and real number $\epsilon > 0$, we have:

$$\sup_{\substack{\pi \in \mathbb{D}(\mathcal{X}): \\ \pi_x > 0}} \max_{x \in \mathcal{X}, y \in Y} \frac{C_{x,y}}{p(y)} = \max_{x,x',y} \frac{C_{x,y}}{C_{x',y}}$$

Proof. First notice that $\frac{C_{x,y}}{C_{x',y}} \leq e^{\epsilon}$ for all $x, x' \in \mathcal{X}$ and $y \in Y$ if and only if the right side of the equation above is also smaller than or equal to e^{ϵ}: if the right side of the equation is smaller than or equal to ϵ, then all pertinent values of x, x', y would lead to values that can not be bigger for $\frac{C_{x,y}}{C_{x',y}}$, so it would also be smaller than or equal to e^{ϵ}. Also, if $\frac{C_{x,y}}{C_{x',y}} \leq e^{\epsilon}$ for all $x, x' \in \mathcal{X}$ and $y \in Y$, then in particular the maximum value of $\frac{C_{x,y}}{C_{x',y}}$ will also be smaller than or equal to e^{ϵ}.

Now let the right-hand side of the equation of this theorem be equal to $k \geq 1$. If we choose ϵ such that $e^{\epsilon} = k$, then the left-hand side of the first equation of Theorem 1 will be smaller than or equal to e^{ϵ}, which we can conclude by the argument of the last paragraph and the appication of Theorem 1. But by the same reasons, (assuming that $k > 1$) any value ϵ' such that $e^{\epsilon'} < k$ will lead to the conclusion that the left-hand side of the first equation of Theorem 1 is bigger than $e^{\epsilon'}$. So, it must be exatly equal to $e^{\epsilon} = k$. If $k = 1$, then all values of $C_{x,y}$ are equal (otherwise we could obtain a fraction bigger than one for entries in the channel), the channel is constant and the result holds. Notice that this argument is only true becaouse of the extra assumption that the channel has no zero entries. □

We did not remove the extra assumption of non-zero entries because this attempt did not work even with that extra assumption, as we will now discuss.

Changing the MAX axiom 11 to the MAX-δ axiom 15 on Theorem 2 and the ϵ-LDP 4 requirement to the (δ, ϵ)-LDP 5 requirement leads to a false logical expression, which we present here as Theorem 3. The goal was, of course, to prove that the expression was true, but we found that this is not the case.

Theorem 3 (Modeling (ϵ, δ)-LDP requires another approach)**.** *The following affirmation is not true.*

For any channel $C : \mathcal{X} \to \mathbb{D}(Y)$ with only non-zero entries and real number $\epsilon > 0$, we have:

$$\sup_{\substack{\pi \in \mathbb{D}(\mathcal{X}): \\ \pi_x > 0}} \max_{\substack{x \in \mathcal{X}, y \in Y: \\ p(y) > \delta}} \frac{C_{x,y}}{p(y)} = \max_{x,x',y} \frac{C_{x,y} - \delta}{C_{x',y}}$$

Proof. Consider the constant channel, in which all entries are equal to $c = \frac{1}{n}$, in which $n = |Y|$. Now, $p(y) = \frac{1}{n}$ regardless of the choice of $\pi \in \mathbb{D}(\mathcal{X})$, so $\frac{C_{x,y}}{p(y)} = \frac{c}{c} = 1$, but $\frac{C_{x,y}-\delta}{C_{x',y}} = 1 - \frac{\delta}{c}$, for any x, x', y. As $\delta > 0$, this equality will be false regardless of the value of c, so the expression presented can not be true for all channels, as it is not true for this specific channel. □

Notice that $\frac{C_{x,y}-\delta}{C_{x',y}} \leq \epsilon \iff P(Y = y|X = x) \leq e^{\epsilon}P(Y = y|X = x') + \delta$, which is equivalent to (ϵ, δ)-LDP 5 when $|S| = 1$. As explained before, we do not attempt to extend the results in this work to more general configurations because the results were already negative in the current, simpler, configuration.

We can conclude that to model $(\delta, \epsilon) - LDP$ in the QIF framework, we can not use the same bound as in the analyzed paper [54]. Other approaches might still be successfull, though, changing or not changing the classical AVG 12 axiom from QIF.

IV. Conclusions and Future Work

We can summarize the results obtained as following:

1) There are methodological shortcomings on some recent studies that show that LDP obfuscation mechanisms can improve fairness of Machine Learning models. In short, the results, in general, depend on trusting the data collector to not try to reverse the noise, and it also limits what trustworthy data collectors can do.
2) The distribution of privacy budget among variables needs to assume some prior knowledge of the data distribution. If no prior knowledge is assumed, applying noise to some but not all variables can not guarantee any level of privacy as the other variables might be used to discover the value of obfuscated variables. Empirically, though, there are results indicating a different noise distribution might be useful [47].
3) In order to provide bounds on (ϵ, δ)-LDP based with the δ-MAX axiom 15 for QIF, bounds different from the ones presented in [54] needs to be explored. Also, even if bounds are found, it would be necessary to prove that the δ-MAX has the useful properties that classical QIF axioms have. We also might be able to model (ϵ, δ)-LDP with the QIF framework via other approaches, not necessarily creating a new QIF axiom.

Possibilities for future work include obtaining theoretical results in respect to the tradeoff between fairness and privacy that take into account the possibility of the data collector trying to reverse the noise. Also, we can model different levels of prior knowledge of a distribution and use them to decide the optimal privacy budget distribution given what the people applying the nosie know a priori. Finally, we can explore another routes to modeling (ϵ, δ)-LDP whith QIF concepts.

References

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

[2] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 60–69. [Online]. Available: https://proceedings.mlr.press/v80/agarwal18a.html

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

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

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

[6] K. Makhlouf, S. Zhioua, and C. Palamidessi, "Survey on causal-based machine learning fairness notions," 2022.

[7] J. Zhou and T. Joachims, "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*, ser. FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 12–21. [Online]. Available: https://doi.org/10.1145/3593013.3593972

[8] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "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.

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

[10] L. Li, T. Lassiter, J. Oh, and M. K. Lee, "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*, 2021, pp. 166–176.

[11] K. Makhlouf, S. Zhioua, and C. Palamidessi, "On the applicability of machine learning fairness notions," *SIGKDD Explor. Newsl.*, vol. 23, no. 1, p. 14–23, may 2021. [Online]. Available: https://doi.org/10.1145/3468507.3468511

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

[13] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "The (im)possibility of fairness: different value systems require different mechanisms for fair decision making," *Commun. ACM*, vol. 64, no. 4, p. 136–143, mar 2021. [Online]. Available: https://doi.org/10.1145/3433949

[14] G. Alves, F. Bernier, M. Couceiro, K. Makhlouf, C. Palamidessi, and S. Zhioua, "Survey on fairness notions and related tensions," *EURO Journal on Decision Processes*, vol. 11, p. 100033, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2193943823000067

[15] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

[16] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–36, 2021.

[17] M. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith, *The Science of Quantitative Information Flow*, ser. Information Security and Cryptography. Springer International

Publishing, 2020. [Online]. Available: https://books.google.com.br/books?id=jJH-DwAAQBAJ

[18] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi, "On the information leakage of differentially-private mechanisms," *Journal of Computer Security*, vol. 23, no. 4, pp. 427–469, 2015.

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

[20] O. Calin, *Deep learning architectures*. Springer, 2020.

[21] P. Grohs and G. Kutyniok, *Mathematical aspects of deep learning*. Cambridge University Press, 2022.

[22] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[23] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[24] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. USA: Cambridge University Press, 2009.

[25] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *arXiv preprint arXiv:1810.08810*, 2018.

[26] G. Alves, F. Bernier, M. Couceiro, K. Makhlouf, C. Palamidessi, and S. Zhioua, "Survey on fairness notions and related tensions," *EURO journal on decision processes*, vol. 11, p. 100033, 2023.

[27] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *International conference on machine learning*. PMLR, 2018, pp. 2564–2572.

[28] C. Dwork and C. Ilvento, "Fairness under composition," *arXiv preprint arXiv:1806.06122*, 2018.

[29] D. Hellman, "Measuring algorithmic fairness," *Virginia Law Review*, vol. 106, no. 4, pp. 811–866, 2020.

[30] A. Bell, L. Bynum, N. Drushchak, T. Zakharchenko, L. Rosenblatt, and J. Stoyanovich, "The possibility of fairness: Revisiting the impossibility theorem in practice," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 400–422.

[31] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[32] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction {APIs}," in *25th USENIX security symposium (USENIX Security 16)*, 2016, pp. 601–618.

[33] B. Wang and N. Z. Gong, "Stealing hyperparameters in machine learning," in *2018 IEEE symposium on security and privacy (SP)*. IEEE, 2018, pp. 36–52.

[34] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[35] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," *International Journal of Security and Networks*, vol. 10, no. 3, pp. 137–150, 2015.

[36] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–37, 2022.

[37] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1897–1914.

[38] X. Xiong, S. Liu, D. Li, Z. Cai, and X. Niu, "A comprehensive survey on local differential privacy," *Security and Communication Networks*, vol. 2020, no. 1, p. 8829523, 2020.

[39] C. Liu, S. Chakraborty, and P. Mittal, "Dependence makes you vulnberable: Differential privacy under dependent tuples." in *NDSS*, vol. 16, 2016, pp. 21–24.

[40] M. C. Tschantz, S. Sen, and A. Datta, "Sok: Differential privacy as a causal property," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 354–371.

[41] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *IEEE INFOCOM 2019-IEEE conference on computer communications*. IEEE, 2019, pp. 2512–2520.

[42] Z. Yang, S. Zhong, and R. N. Wright, "Privacy-preserving classification of customer data without loss of accuracy," in *Proceedings of the 2005 SIAM International Conference on Data Mining*. SIAM, 2005, pp. 92–102.

[43] E. Hesamifard, H. Takabi, and M. Ghasemi, "Cryptodl: Deep neural networks over encrypted data," *arXiv preprint arXiv:1711.05189*, 2017.

[44] M. J. Goswami, "Privacy-preserving deep learning using secure multi-party computation," *International IT Journal of Research, ISSN: 3007-6706*, vol. 2, no. 2, pp. 50–55, 2024.

[45] K. Makhlouf, T. Stefanovic, H. H. Arcolezi, and C. Palamidessi, "A systematic and formal study of the impact of local differential privacy on fairness: Preliminary results," 2024. [Online]. Available: https://arxiv.org/abs/2405.14725

[46] K. Makhlouf, H. H. Arcolezi, S. Zhioua, G. B. Brahim, and C. Palamidessi, "On the impact of multi-dimensional local differential privacy on fairness," *Data Mining and Knowledge Discovery*, pp. 1–24, 2024.

[47] H. H. Arcolezi, K. Makhlouf, and C. Palamidessi, "(local) differential privacy has no disparate impact on fairness," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2023, pp. 3–21.

[48] C. P. Henao, "Exploring fairness and privacy in machine learning," Ph.D. dissertation, Institut Polytechnique de Paris, 2023.

[49] D. Franco, L. Oneto, N. Navarin, and D. Anguita, "Toward learning trustworthily from data combining privacy, fairness, and explainability: an application to face recognition," *Entropy*, vol. 23, no. 8, p. 1047, 2021.

[50] M. Alvim, N. Fernandes, B. Nogueira, C. Palamidessi, and T. Silva, "On the duality of privacy and fairness (extended abstract)," in *International Conference on AI and the Digital Economy (CADE 2023)*. United Kingdom: Institution of Engineering and Technology, 2023, 9th International Conference on AI and the Digital Economy, CADE 2023 ; Conference date: 26-06-2023 Through 28-06-2023.

[51] B. D. Nogueira *et al.*, "On the relation of privacy and fairness through the lenses of quantitative information flow," 2023.

[52] X. Ren, C.-M. Yu, W. Yu, S. Yang, X. Yang, J. A. McCann, and S. Y. Philip, "Lopub: high-dimensional crowdsourced data publication with local differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2151–2166, 2018.

[53] N. Konstantinov and C. H. Lampert, "On the impossibility of fairness-aware learning from corrupted data," in *Algorithmic Fairness through the Lens of Causality and Robustness workshop*. PMLR, 2022, pp. 59–83.

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