On the Impact of Local Differential Privacy on Fairness: A Formal Approach

Karima Makhlouf

Inria and École Polytechnique

A joint work with:

Tamara Stefanović, Héber H. Arcolezi, and Catuscia Palamidessi

8th Franco-Japanese Cybersecurity Workshop

November 29, 2023

Overview

Overview

Motivation

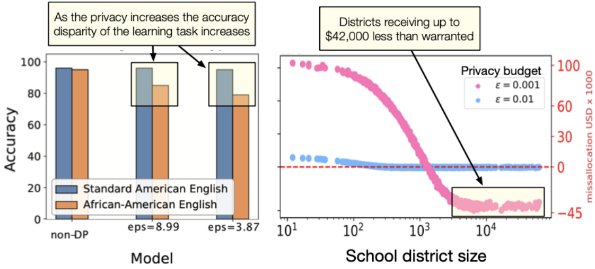


Figure 1: Impact of Differential Privacy on Fairness. Image from [1]

[1] Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy and fairness in decisions and learning tasks:A survey (2022).

Motivation

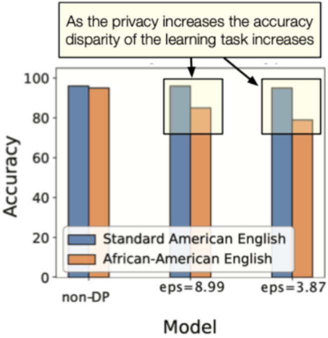


Figure 2: Impact of Differential Privacy on Fairness. Image from [1]

[1] Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy and fairness in decisions and learning tasks:A survey (2022).

Motivation

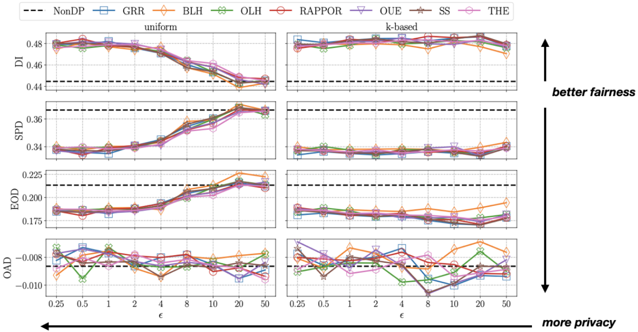
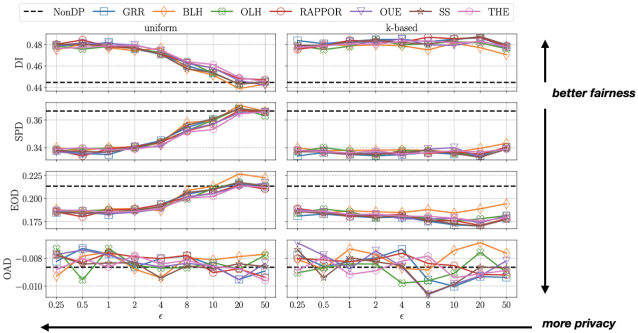


Figure 3: Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (uniform on the left-side and our k-based on the right-side), on the Adult dataset [2].

[2] Héber H. Arcolezi, Karima Makhlouf, and Catuscia Palamidessi. (local) differential privacy has NO disparate impact on fairness. In Data and Applications Security and Privacy XXXVII, pages 3–21. Springer Nature Switzerland, 2023.

Motivation



Fairness issues in DP settings are receiving increasing attention

BUT

complete understanding of why is not well explored!

Overview

Fairness

Informal definition

Absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making [3].

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. ACM Computing Surveys, 2021.

Fairness

Informal definition

Absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making [3].

Data

$A \in \{0,1\}, X \in dom(X)$: sensitive attribute, non-sensitive attributes
$Y, \hat{Y} \in \{0,1\}$: true decision, prediction of the classifier

Fairness metric	Abbriv.	Formula
Statistical Disparity	SD	$\mathbb{P}[\hat{Y}=1 \mid A=1] - \mathbb{P}[\hat{Y}=1 \mid A=0]$
Conditional Statistical Disparity	CSD_x	$\mathbb{P}[\hat{Y}=1 \mid X=x, A=1] - \mathbb{P}[\hat{Y}=0 \mid X=x, A=0]$
Equal Opportunity Disparity	EOD	$\mathbb{P}[\hat{Y}=1 \mid Y=1, A=1] - \mathbb{P}[\hat{Y}=1 \mid Y=1, A=0]$
Predictive Equality Disparity	PED	$\mathbb{P}[\hat{Y}=1 \mid Y=0, A=1] - \mathbb{P}[\hat{Y}=1 \mid Y=0, A=0]$
Overall Accuracy Disparity	OAD	$\mathbb{P}[\hat{Y}=Y \mid A=1] - \mathbb{P}[\hat{Y}=Y \mid A=0]$

Table 1: Some fairness metrics.

Fairness

Informal definition

Absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making [3].

Data

$A \in \{0,1\}, X \in dom(X)$: sensitive attribute, non-sensitive attributes
$Y, \hat{Y} \in \{0,1\}$: true decision, prediction of the classifier

Fairness metric	Abbriv.	Formula
Statistical Disparity	**SD**	$\mathbb{P}[\hat{\mathbf{Y}} = \mathbf{1} \mid \mathbf{A} = \mathbf{1}] - \mathbb{P}[\hat{\mathbf{Y}} = \mathbf{1} \mid \mathbf{A} = \mathbf{0}]$
Conditional Statistical Disparity	$\mathbf{CSD_x}$	$\mathbb{P}[\hat{\mathbf{Y}} = \mathbf{1} \mid \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{1}] - \mathbb{P}[\hat{\mathbf{Y}} = \mathbf{0} \mid \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{0}]$
Equal Opportunity Disparity	*EOD*	$\mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 1] - \mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 0]$
Predictive Equality Disparity	*PED*	$\mathbb{P}[\hat{Y} = 1 \mid Y = 0, A = 1] - \mathbb{P}[\hat{Y} = 1 \mid Y = 0, A = 0]$
Overall Accuracy Disparity	*OAD*	$\mathbb{P}[\hat{Y} = Y \vert A = 1] - \mathbb{P}[\hat{Y} = Y \vert A = 0]$

Table 2: Some fairness metrics.

Local Differential Privacy (LDP)

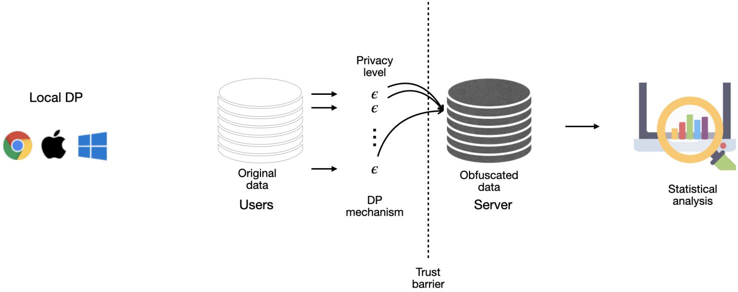


Figure 4: Local differential privacy.

Local Differential Privacy (LDP)

Definition ($\epsilon - LDP$).

An algorithm $\mathcal{M}$ satisfies ϵ-local-differential-privacy (ϵ-LDP), where $\epsilon > 0$, if for any input v_1 and $v_2 \in Dom(\mathcal{M})$ and $\forall$ possible output $y \in Dom(\mathcal{M})$ [3]:

$$\mathbb{P}[\mathcal{M}(v_1) = y] \leq e^{\epsilon} \, \mathbb{P}[\mathcal{M}(v_2) = y]$$

[3] Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately? In: 2008 49th Annual IEEE Symposium on Foundations of Computer Science.

Overview

Problem Definition

- Study formally the impact of LDP on fairness.
  - Quantify the impact of LDP on the disparity between groups (e.g., CSD_x, SD, etc.).
  - Provide bounds in terms of the joint distributions and the privacy level, delimiting the extent by which LDP can impact fairness.

Problem Definition

- Study formally the impact of LDP on fairness.
  - Quantify the impact of LDP on the disparity between groups (e.g., CSD_x, SD, etc.).
  - Provide bounds in terms of the joint distributions and the privacy level, delimiting the extent by which LDP can impact fairness.
- Validate our theoretical findings empirically with synthetic and real-world datasets.

Problem Definition

- Study formally the impact of LDP on fairness.
  - Quantify the impact of LDP on the disparity between groups (e.g., CSD_x, SD, etc.).
  - Provide bounds in terms of the joint distributions and the privacy level, delimiting the extent by which LDP can impact fairness.
- Validate our theoretical findings empirically with synthetic and real-world datasets.

Note: We apply privacy only to $A \to A' \quad (A' = \mathcal{M}(A))$
$\hat{Y} \to \hat{Y}'$

Problem Definition

- Study formally the impact of LDP on fairness.
  - Quantify the impact of LDP on the disparity between groups (e.g., CSD_x, SD, etc.).
  - Provide bounds in terms of the joint distributions and the privacy level, delimiting the extent by which LDP can impact fairness.
- Validate our theoretical findings empirically with synthetic and real-world datasets.

Note: We apply privacy only to $A \rightarrow A'$ $(A' = \mathcal{M}(A))$
$\hat{Y} \rightarrow \hat{Y}'$

- LDP mechanism

$$\mathcal{M}(a) = \begin{cases} a & \text{with} \quad p, \\ \overline{a} & \text{with} \quad 1-p. \end{cases} \qquad \text{where } p = \frac{e^{\epsilon}}{e^{\epsilon}+1}$$

$$\frac{p}{1-p} = e^{\epsilon}$$

Overview

Theoretical Results: Notations and Definitions

Data

$A, A' \in \{0,1\}$: sensitive attribute before obfuscation, after obfuscation

$X \in dom(X)$: non-sensitive attributes

$Y \in \{0,1\}$: true decision

$\hat{Y}, \hat{Y}' \in \{0,1\}$: prediction of the classifier before obfuscation, after obfuscation

Theoretical Results: Notations and Definitions

Data

$A, A' \in \{0, 1\}$: sensitive attribute before obfuscation, after obfuscation

$X \in dom(X)$: non-sensitive attributes

$Y \in \{0, 1\}$: true decision

$\hat{Y}, \hat{Y}' \in \{0, 1\}$: prediction of the classifier before obfuscation, after obfuscation

Definitions

- $\Gamma_a^x = \hat{\mathbb{P}}[Y = 1|X = x, A = a] - \hat{\mathbb{P}}[Y = 0|X = x, A = a]$
- $\Delta_a^x = \hat{\mathbb{P}}[Y = 1, X = x, A = a] - \hat{\mathbb{P}}[Y = 0, X = x, A = a]$
- $\Gamma_a'^x = \hat{\mathbb{P}}[Y = 1|X = x, A' = a] - \hat{\mathbb{P}}[Y = 0|X = x, A' = a]$
- $\Delta_a'^x = \hat{\mathbb{P}}[Y = 1, X = x, A' = a] - \hat{\mathbb{P}}[Y = 0, X = x, A' = a]$

Theoretical Results: Assumptions

- ML model (baseline)

$$\mathbb{P}[\hat{Y} = 1 | X = x, A = a] = \hat{Y}_a^x = \begin{cases} 1 & \text{if} \quad \Delta_a^x \geq 0 \quad \text{(equiv. } \Gamma_a^x \geq 0\text{)}, \\ 0 & \text{otherwise.} \end{cases}$$

Theoretical Results: Assumptions

- ML model (baseline)

$$\mathbb{P}[\hat{Y} = 1 | X = x, A = a] = \hat{Y}_a^x = \begin{cases} 1 & \text{if} \quad \Delta_a^x \geq 0 \quad (\text{equiv. } \Gamma_a^x \geq 0), \\ 0 & \text{otherwise.} \end{cases}$$

- ML model (after obfuscation)

$$\mathbb{P}[\hat{Y} = 1 | X = x, A' = a] = \hat{Y}_a'^x = \begin{cases} 1 & \text{if} \quad \Delta_a'^x \geq 0 \quad (\text{equiv. } \Gamma_a'^x \geq 0), \\ 0 & \text{otherwise.} \end{cases}$$

Theoretical Results

Lemma 1

$\Delta_a'^x = p\ \Delta_a^x + (1-p)\ \Delta_{\bar{a}}^x$

Theoretical Results

Lemma 1

$\Delta_a'^x = p\ \Delta_a^x + (1-p)\ \Delta_{\bar{a}}^x$

Lemma 2

$$
\begin{aligned}
\text{- } \hat{Y}_a'^x = 1 \quad & \text{if} \quad \Delta_a^x, \Delta_{\bar{a}}^x \geq 0 \\
& \text{or} \quad \Delta_a^x > 0 \quad \text{and } \Delta_{\bar{a}}^x < 0 \quad \text{and} \quad e^\epsilon \geq -\frac{\Delta_{\bar{a}}^x}{\Delta_a^x} \\
& \text{or} \quad \Delta_a^x < 0 \quad \text{and } \Delta_{\bar{a}}^x > 0 \quad \text{and} \quad e^\epsilon \leq -\frac{\Delta_{\bar{a}}^x}{\Delta_a^x}
\end{aligned}
$$

Theoretical Results

Lemma 1

$\Delta_a'^{x} = p\ \Delta_a^x + (1-p)\ \Delta_{\bar{a}}^x$

Lemma 2

$$
\begin{aligned}
\text{- } \hat{Y}_a'^{x} = 1 \quad & \text{if} \quad \Delta_a^x, \Delta_{\bar{a}}^x \geq 0 \\
& \text{or} \quad \Delta_a^x > 0 \quad \text{and } \Delta_{\bar{a}}^x < 0 \quad \text{and} \quad e^\epsilon \geq -\frac{\Delta_{\bar{a}}^x}{\Delta_a^x} \\
& \text{or} \quad \Delta_a^x < 0 \quad \text{and } \Delta_{\bar{a}}^x > 0 \quad \text{and} \quad e^\epsilon \leq -\frac{\Delta_{\bar{a}}^x}{\Delta_a^x} \\
\text{- } \hat{Y}_a'^{x} = 0 \quad & \text{if} \quad \Delta_a^x, \Delta_{\bar{a}}^x \leq 0 \quad \text{and at least one of them} < 0 \\
& \text{or} \quad \Delta_a^x > 0 \quad \text{and } \Delta_{\bar{a}}^x < 0 \quad \text{and} \quad e^\epsilon < -\frac{\Delta_{\bar{a}}^x}{\Delta_a^x} \\
& \text{or} \quad \Delta_a^x < 0 \quad \text{and } \Delta_{\bar{a}}^x > 0 \quad \text{and} \quad e^\epsilon > -\frac{\Delta_{\bar{a}}^x}{\Delta_a^x}
\end{aligned}
$$

Impact of LDP on $\mathsf{CSD_x}$

Theoretical Results for CSD_x

Reminder: $\hat{Y}_a^x = \mathbb{P}[\hat{Y} = 1 | X = x, A = a]$

Definition (CSD_x)

$CSD_x \stackrel{def}{=} \hat{Y}_1^x - \hat{Y}_0^x$

Definition (CSD'_x)

$CSD'_x \stackrel{def}{=} \hat{Y}'^x_1 - \hat{Y}'^x_0$

Theoretical Results for CSD_x

Reminder: $\hat{Y}_a^x = \mathbb{P}[\hat{Y} = 1 | X = x, A = a]$

Definition (CSD_x)

$CSD_x \overset{def}{=} \hat{Y}_1^x - \hat{Y}_0^x$

Definition (CSD'_x)

$CSD'_x \overset{def}{=} \hat{Y'}_1^x - \hat{Y'}_0^x$

Theorem (Impact of LDP on CSD_x)

1. if $CSD_x > 0$ then $0 \leq CSD'_x \leq CSD_x$
2. if $CSD_x < 0$ then $CSD_x \leq CSD'_x \leq 0$
3. if $CSD_x = 0$ then $CSD'_x = CSD_x = 0$

Impact of LDP on SD

$$(X \perp A)$$

Theoretical Results for *SD*

Definition (*SD*)

$$SD \stackrel{def}{=} \mathbb{P}[\hat{Y} = 1|A = 1] - \mathbb{P}[\hat{Y} = 1|A = 0]$$

Definition (*SD'*)

$$SD' \stackrel{def}{=} \mathbb{P}[\hat{Y}' = 1|A = 1] - \mathbb{P}[\hat{Y}' = 1|A = 0]$$

Theoretical Results for SD $(X \perp A)$

Uniformity Assumption

if $\exists x^* : \Gamma_a^{x^*} > \Gamma_{\bar{a}}^{x^*}$ then $\forall x \;\; \Gamma_a^{x} \geq \Gamma_{\bar{a}}^{x}$

Theoretical Results for SD ($X \perp A$)

Uniformity Assumption

if $\exists x^* : \Gamma_a^{x^*} > \Gamma_{\bar{a}}^{x^*}$ then $\forall x \;\; \Gamma_a^x \geq \Gamma_{\bar{a}}^x$

Lemma 3

$$SD = \begin{cases} \mathbb{P}[\Delta_1^X \geq 0 \;\wedge\; \Delta_0^X < 0] & \text{if } \exists x \; \Gamma_1^x > \Gamma_0^x \\ 0 & \text{if } \forall x \; \Gamma_1^x = \Gamma_0^x \\ -\,\mathbb{P}[\Delta_1^X < 0 \;\wedge\; \Delta_0^X \geq 0] & \text{if } \exists x \; \Gamma_1^x < \Gamma_0^x \end{cases}$$

Theoretical Results for SD ($X \perp A$)

Lemma 4

$$SD' = \begin{cases} \mathbb{P}[\Delta_1'^X \geq 0 \;\wedge\; \Delta_0'^X < 0] & \text{if } \exists x \; \Gamma_1'^x > \Gamma_0'^x \\ 0 & \text{if } \forall x \; \Gamma_1'^x = \Gamma_0'^x \\ -\mathbb{P}[\Delta_1'^X < 0 \;\wedge\; \Delta_0'^X \geq 0] & \text{if } \exists x \; \Gamma_1'^x < \Gamma_0'^x \end{cases}$$

$$SD' = \begin{cases} \mathbb{P}[\Delta_1^X > 0 \;\wedge\; \Delta_0^X < 0 \;\wedge\; e^\epsilon \geq -\frac{\Delta_0^X}{\Delta_1^X} \;\wedge\; e^\epsilon > -\frac{\Delta_1^X}{\Delta_0^X}] & \text{if } \exists x \; \Gamma_1^x > \Gamma_0^x \\ 0 & \text{if } \forall x \; \Gamma_1^x = \Gamma_0^x \\ -\mathbb{P}[\Delta_1^X < 0 \;\wedge\; \Delta_0^X > 0 \;\wedge\; e^\epsilon > -\frac{\Delta_0^X}{\Delta_1^X} \;\wedge\; e^\epsilon \geq -\frac{\Delta_1^X}{\Delta_0^X}] & \text{if } \exists x \; \Gamma_1^x < \Gamma_0^x \end{cases}$$

Note: If ϵ is small enough (i.e., $\forall x \;\; e^\epsilon < -\frac{\Delta_0^x}{\Delta_1^x}$ or $e^\epsilon < -\frac{\Delta_1^x}{\Delta_0^x}$) $\rightarrow SD' = 0$.

Theoretical Results for SD $(X \perp A)$

Theorem (Impact of LDP on SD $(X \perp A)$)

1. if $SD > 0$ then $0 \leq SD' \leq SD$
2. if $SD < 0$ then $SD \leq SD' \leq 0$
3. if $SD = 0$ then $SD' = SD = 0$

Impact of LDP on SD

$$(X \not\perp A)$$

Theoretical Results for SD ($X \not\perp A$)

Theorem (Impact of LDP on SD ($X \not\perp A$))

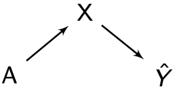
1. if $\exists\ x\ \Gamma_1^x > \Gamma_0^x$ then $SD' \leq SD$
2. if $\exists\ x\ \Gamma_1^x < \Gamma_0^x$ then $SD \leq SD'$
3. if $\forall\ x\ \Gamma_1^x = \Gamma_0^x$ then $SD' = SD$

Notes

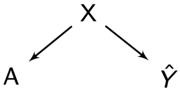
- SD' and SD may have opposite signs.
- In (1), we could have $SD < 0$ ($\mathbb{P}[X = x|A = 1] \ll [X = x|A = 0]$) $\rightarrow$ *Simpson paradox*.
- Similarly, for case (2), we could have $SD > 0$.
- In general, the unprivileged group benefits from LDP.

Overview

Some causality



(a) A mediator structure



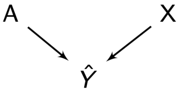
(b) A confounder structure

$$A \not\perp \hat{Y}$$

$$A \perp \hat{Y} | X$$

$$CSD'_x = CSD_x = 0$$

$$SD' = SD$$



(c) A collider structure

$$A \perp X$$

$$A \not\perp X | \hat{Y}$$

$$0 \leq SD' \leq SD$$

$$SD \leq SD' \leq 0$$

Overview

Takeaways and Future directions

- Privacy and fairness can go hand in hand (decreasing disparity between groups)
- In general, the unprivileged group benefits from privacy
- Privacy does not bring fake discrimination

- Expand our study to other fairness notions (EOD, OAD, etc.)
- Considering LDP multi-dimensional data (we have some preliminary empirical results on synthetic and real-world dataset)
- Considering more in-depth causality (confounders, mediators, colliders)

Thanks