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

November 23, 2023
Overview

1. Motivation
2. Background about Fairness and LDP
3. Problem Definition
4. Theoretical Results
5. Some Causality
6. Takeaways and Future directions
Overview

1. Motivation

2. Background about Fairness and LDP

3. Problem Definition

4. Theoretical Results

5. Some Causality

6. Takeaways and Future directions
Motivation

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

Motivation

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

Figure 3: Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the $\epsilon$-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].

Fairness issues in DP settings are receiving increasing attention

**BUT**

complete understanding of why is not well explored!
Overview

1. Motivation

2. Background about Fairness and LDP

3. Problem Definition

4. Theoretical Results

5. Some Causality

6. Takeaways and Future directions
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].

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

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

<table>
<thead>
<tr>
<th>Fairness metric</th>
<th>Abbriv.</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Disparity</td>
<td>SD</td>
<td>(P[\hat{Y} = 1</td>
</tr>
<tr>
<td>Conditional Statistical Disparity</td>
<td>CSD(_x)</td>
<td>(P[\hat{Y} = 1</td>
</tr>
<tr>
<td>Equal Opportunity Disparity</td>
<td>EOD</td>
<td>(P[\hat{Y} = 1</td>
</tr>
<tr>
<td>Predictive Equality Disparity</td>
<td>PED</td>
<td>(P[\hat{Y} = 1</td>
</tr>
<tr>
<td>Overall Accuracy Disparity</td>
<td>OAD</td>
<td>(P[\hat{Y} = Y</td>
</tr>
</tbody>
</table>

**Table 1:** Some fairness metrics.
**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**

\(A \in \{0, 1\}, X \in \text{dom}(X)\): sensitive attribute, non-sensitive attributes

\(Y, \hat{Y} \in \{0, 1\}\): true decision, prediction of the classifier

<table>
<thead>
<tr>
<th>Fairness metric</th>
<th>Abbrev.</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Disparity</td>
<td>SD</td>
<td>(P[\hat{Y} = 1</td>
</tr>
<tr>
<td>Conditional Statistical Disparity</td>
<td>CSD(_x)</td>
<td>(P[\hat{Y} = 1</td>
</tr>
<tr>
<td>Equal Opportunity Disparity</td>
<td>EOD</td>
<td>(P[\hat{Y} = 1</td>
</tr>
<tr>
<td>Predictive Equality Disparity</td>
<td>PED</td>
<td>(P[\hat{Y} = 1</td>
</tr>
<tr>
<td>Overall Accuracy Disparity</td>
<td>OAD</td>
<td>(P[\hat{Y} = Y</td>
</tr>
</tbody>
</table>

**Table 2:** Some fairness metrics.
Figure 4: Local differential privacy.
Local Differential Privacy (LDP)

**Definition ($\epsilon - LDP$).**

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

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

Overview

1. Motivation
2. Background about Fairness and LDP
3. Problem Definition
4. Theoretical Results
5. Some Causality
6. Takeaways and Future directions
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 $A \rightarrow A'$ ($A' = \mathcal{M}(A)$)
\[\hat{Y} \rightarrow \hat{Y}'\]
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

\[
\mathcal{M}(a) = \begin{cases} 
    a & \text{with } p, \\
    \overline{a} & \text{with } 1 - p.
\end{cases}
\]

where \( p = \frac{e^\epsilon}{e^\epsilon + 1} \)

\[
\frac{p}{1-p} = e^\epsilon
\]
Overview

1 Motivation

2 Background about Fairness and LDP

3 Problem Definition

4 Theoretical Results

5 Some Causality

6 Takeaways and Future directions
Data

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

$X \in \text{dom}(X)$: 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
- \( X \in \text{dom}(X) \): 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

- \( \Gamma^x_a = \hat{P}[Y = 1|X = x, A = a] - \hat{P}[Y = 0|X = x, A = a] \)
- \( \Delta^x_a = \hat{P}[Y = 1, X = x, A = a] - \hat{P}[Y = 0, X = x, A = a] \)
- \( \Gamma'^{x'}_a = \hat{P}[Y = 1|X = x, A' = a] - \hat{P}[Y = 0|X = x, A' = a] \)
- \( \Delta'^{x'}_a = \hat{P}[Y = 1, X = x, A' = a] - \hat{P}[Y = 0, X = x, A' = a] \)
Theoretical Results: Assumptions

- ML model (baseline)

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

- ML model (baseline)

\[ P[\hat{Y} = 1|X = x, A = a] = \hat{Y}^x_a = \begin{cases} 1 & \text{if } \Delta^x_a \geq 0 \quad (\text{equiv. } \Gamma^x_a \geq 0), \\ 0 & \text{otherwise.} \end{cases} \]

- ML model (after obfuscation)

\[ P[\hat{Y} = 1|X = x, A' = a] = \hat{Y}'^x_a = \begin{cases} 1 & \text{if } \Delta'^x_a \geq 0 \quad (\text{equiv. } \Gamma'^x_a \geq 0), \\ 0 & \text{otherwise.} \end{cases} \]
Theoretical Results

Lemma 1

\[ \Delta'_a = p \Delta_a^x + (1 - p) \Delta_{\bar{a}}^x \]
Theoretical Results

Lemma 1
\[ \Delta'_a = p \Delta_a + (1 - p) \Delta_a \]

Lemma 2
\[ \hat{Y}'_{a} = 1 \quad \text{if} \quad \Delta_a, \Delta_a \geq 0 \]
\[ \text{or} \quad \Delta_a > 0 \quad \text{and} \quad \Delta_a < 0 \quad \text{and} \quad e^c \geq -\frac{\Delta_a}{\Delta_a} \]
\[ \text{or} \quad \Delta_a < 0 \quad \text{and} \quad \Delta_a > 0 \quad \text{and} \quad e^c \leq -\frac{\Delta_a}{\Delta_a} \]
### Theoretical Results

#### Lemma 1

\[
\Delta_a^x = p \Delta_a^x + (1 - p) \Delta_a^x
\]

#### Lemma 2

- \( \hat{Y}_a^x = 1 \) if \( \Delta_a^x, \Delta_a^x \geq 0 \)
  - or \( \Delta_a^x > 0 \) and \( \Delta_a^x < 0 \) and \( e^\epsilon \geq -\frac{\Delta_a^x}{\Delta_a^x} \)
  - or \( \Delta_a^x < 0 \) and \( \Delta_a^x > 0 \) and \( e^\epsilon \leq -\frac{\Delta_a^x}{\Delta_a^x} \)

- \( \hat{Y}_a^x = 0 \) if \( \Delta_a^x, \Delta_a^x \leq 0 \) and at least one of them \( < 0 \)
  - or \( \Delta_a^x > 0 \) and \( \Delta_a^x < 0 \) and \( e^\epsilon < -\frac{\Delta_a^x}{\Delta_a^x} \)
  - or \( \Delta_a^x < 0 \) and \( \Delta_a^x > 0 \) and \( e^\epsilon > -\frac{\Delta_a^x}{\Delta_a^x} \)
Impact of LDP on $\text{CSD}_x$
Theoretical Results: Results for $CSD_x$

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

**Definition ($CSD_x$)**

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

**Definition ($CSD'_x$)**

$$CSD'_x \overset{\text{def}}{=} \hat{Y}'_1^x - \hat{Y}'_0^x$$
Theoretical Results: Results for $CSD_x$

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

### Definition ($CSD_x$)

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

### Definition ($CSD'_x$)

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

### 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: Results for $SD$

**Definition ($SD$)**

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

**Definition ($SD'$)**

$$SD' \overset{def}{=} P[\hat{Y}' = 1|A = 1] - P[\hat{Y}' = 1|A = 0]$$
Theoretical Results: Results for $SD \ (X \perp A)$

Uniformity Assumption

if $\exists x^*: \Gamma_a^{x^*} > \Gamma_a^{x^*}$ then $\forall x \ \Gamma_a^x \geq \Gamma_a^x$
Theoretical Results: Results for SD ($X \perp A$)

**Uniformity Assumption**

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

**Lemma 3**

$$SD = \begin{cases} 
\mathbb{P}[\Delta_1^X \geq 0 \land \Delta_0^X < 0] & \text{if } \exists x \Gamma_1^x > \Gamma_0^x \\
0 & \text{if } \forall x \Gamma_1^x = \Gamma_0^x \\
-\mathbb{P}[\Delta_1^X < 0 \land \Delta_0^X \geq 0] & \text{if } \exists x \Gamma_1^x < \Gamma_0^x 
\end{cases}$$
Theoretical Results: Results for SD ($X \perp A$)

**Lemma 4**

\[ SD' = \begin{cases} 
\mathbb{P}[\Delta'_1^X \geq 0 \land \Delta'_0^X < 0] & \text{if } \exists x \Gamma'_1^x > \Gamma'_0^x \\
0 & \text{if } \forall x \Gamma'_1^x = \Gamma'_0^x \\
-\mathbb{P}[\Delta'_1^X < 0 \land \Delta'_0^X \geq 0] & \text{if } \exists x \Gamma'_1^x < \Gamma'_0^x 
\end{cases} \]

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

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

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

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

1. if $\exists x \Gamma^x_1 > \Gamma^x_0$ then $SD' \leq SD$
2. if $\exists x \Gamma^x_1 < \Gamma^x_0$ then $SD \leq SD'$
3. if $\forall x \Gamma^x_1 = \Gamma^x_0$ then $SD' = SD = 0$

**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

1. Motivation

2. Background about Fairness and LDP

3. Problem Definition

4. Theoretical Results

5. Some Causality

6. Takeaways and Future directions
Some causality

(a) A mediator structure
(b) A confounder structure
(c) A collider structure

\[ A \not\perp \hat{Y} \]
\[ A \perp \hat{Y} | X \]
\[ CSD'_x = CSD_x = 0 \]
\[ SD' = SD = 0 \]
\[ A \perp X \]
\[ A \not\perp X | \hat{Y} \]
\[ 0 \leq SD' \leq SD \]
\[ SD \leq SD' \leq 0 \]
Overview

1 Motivation

2 Background about Fairness and LDP

3 Problem Definition

4 Theoretical Results

5 Some Causality

6 Takeaways and Future directions
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