(Local) Differential Privacy has NO Disparate Impact on Fairness

Héber H. Arcolezi, Karima Makhlouf, and Catuscia Palamidessi

Inria and École Polytechnique (IPP), Palaiseau, France
{heber.hwang-arcolezi,karima.makhlouf,catuscia}@lix.polytechnique.fr

DBSec, July 19th, 2023

Motivation

Differential Privacy (DP) and Fairness: Friends or Foes?

Fairness Through Awareness

Cynthia Dwork
Microsoft Research S.V.
Mountain View, CA, USA
dwork@microsoft.com

Moritz Hardt*
IBM Research Almaden
San Jose, CA, USA
mhardt@us.ibm.com

Toniann Pitassi†
University of Toronto
Dept. of Computer Science
Toronto, ON, CANADA
toni@cs.toronto.edu

Omer Reingold
Microsoft Research S. V.
Mountain View, CA, USA
omer.reingold@microsoft.com

Richard Zemel‡
University of Toronto
Dept. of Computer Science
Toronto, ON, CANADA
zemel@cs.toronto.edu

An Empirical Analysis of Fairness Notions under Differential Privacy*

Anderson Santana de Oliveira,[1] Caelin Kaplan, [2] Khawla Mallat [1] Tanmay Chakraborty [3]
[1] SAP
[2] SAP and INRIA
[3] SAP and Eurecom
firstname.lastname@sap.com

ON THE APPLICATION AND IMPACT OF ϵ-DP AND FAIRNESS IN AMBULANCE ENGAGEMENT TIME PREDICTION

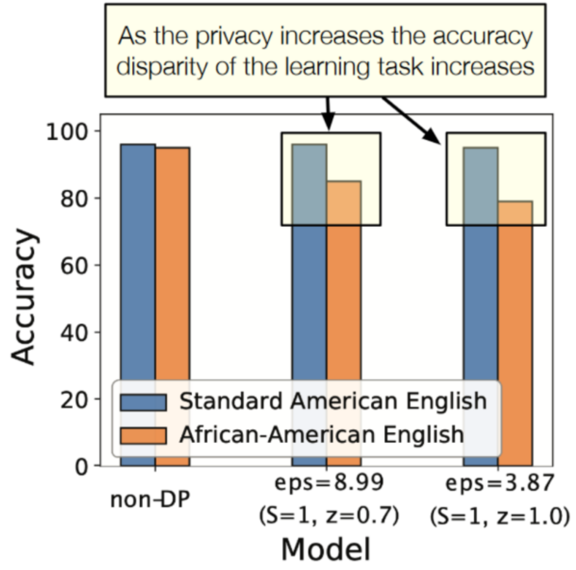
Selene Cerna & Catuscia Palamidessi
Inria and École Polytechnique (IPP), Palaiseau, France
{selene-leya.cerna-nahuis,catuscia.palamidessi}@inria.fr

Differential Privacy Has Disparate Impact on Model Accuracy

Eugene Bagdasaryan
Cornell Tech
eugene@cs.cornell.edu

Omid Poursaeed*
Cornell Tech
op63@cornell.edu

Vitaly Shmatikov
Cornell Tech
shmat@cs.cornell.edu



Differential Privacy (DP) and Fairness: Friends or Foes?

Fairness Through Awareness

Cynthia Dwork
Microsoft Research S.V.
Mountain View, CA, USA
dwork@microsoft.com

Moritz Hardt*
IBM Research Almaden
San Jose, CA, USA
mhardt@us.ibm.com

Toniann Pitassi†
University of Toronto
Dept. of Computer Science
Toronto, ON, CANADA
toni@cs.toronto.edu

Omer Reingold
Microsoft Research S. V.
Mountain View, CA, USA
omer.reingold@microsoft.com

Richard Zemel‡
University of Toronto
Dept. of Computer Science
Toronto, ON, CANADA
zemel@cs.toronto.edu

An Empirical Analysis of Fairness Notions under Differential Privacy*

Anderson Santana de Oliveira,[1] Caelin Kaplan, [2] Khawla Mallat [1] Tanmay Chakraborty [3]

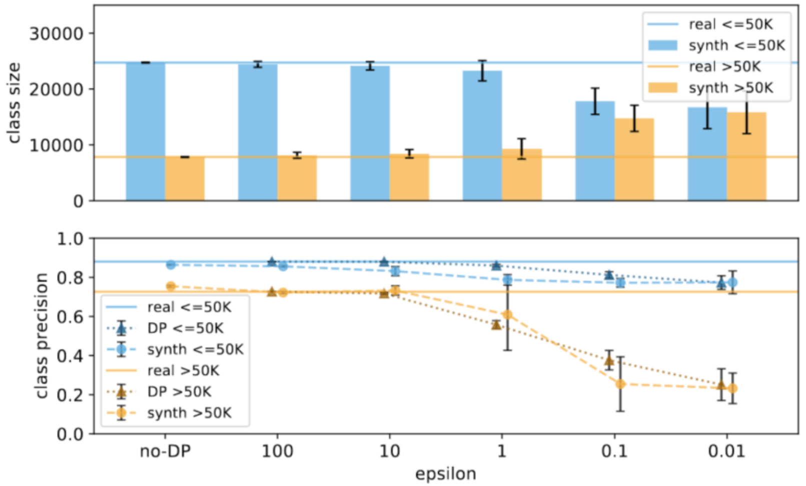
[1] SAP
[2] SAP and INRIA
[3] SAP and Eurecom
firstname.lastname@sap.com

ON THE APPLICATION AND IMPACT OF ϵ-DP AND FAIRNESS IN AMBULANCE ENGAGEMENT TIME PREDICTION

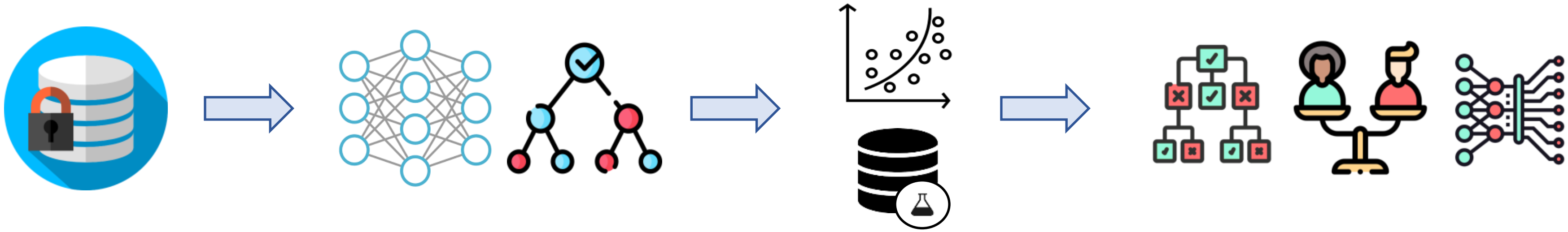
Selene Cerna & Catuscia Palamidessi
Inria and École Polytechnique (IPP), Palaiseau, France
{selene-leya.cerna-nahuis,catuscia.palamidessi}@inria.fr

Robin Hood and *Matthew* Effects: Differential Privacy Has Disparate Impact on Synthetic Data

Georgi Ganev [1 2] Bristena Oprisanu [1] Emiliano De Cristofaro [1]

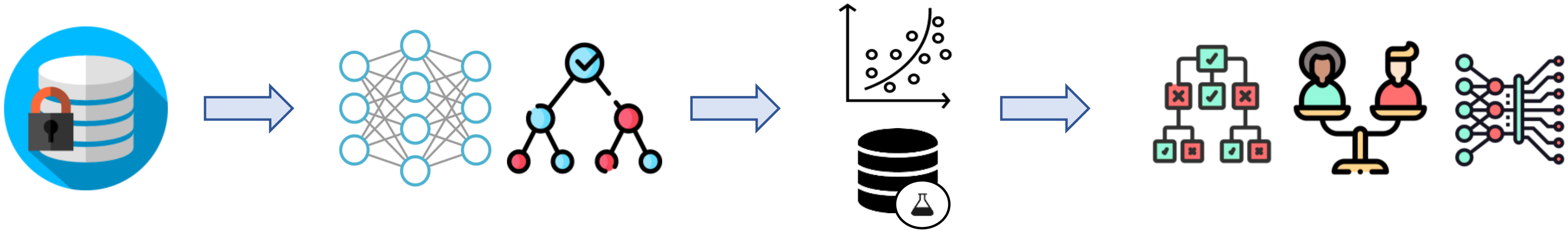


Local DP (LDP) and Fairness: Friends or Foes?

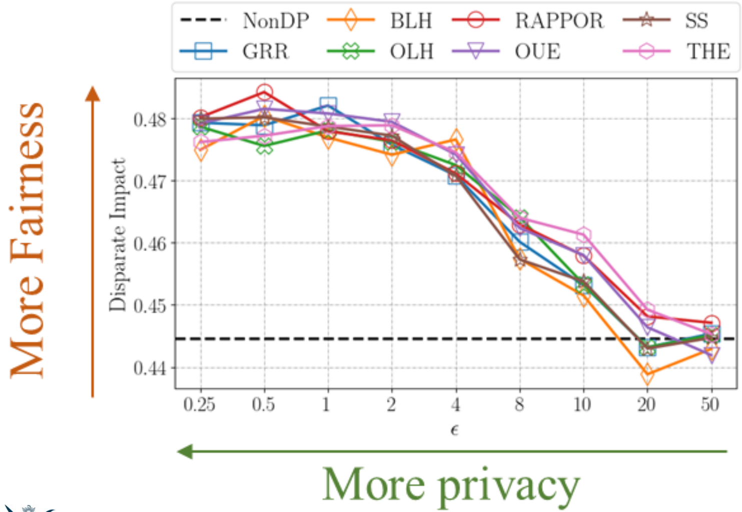


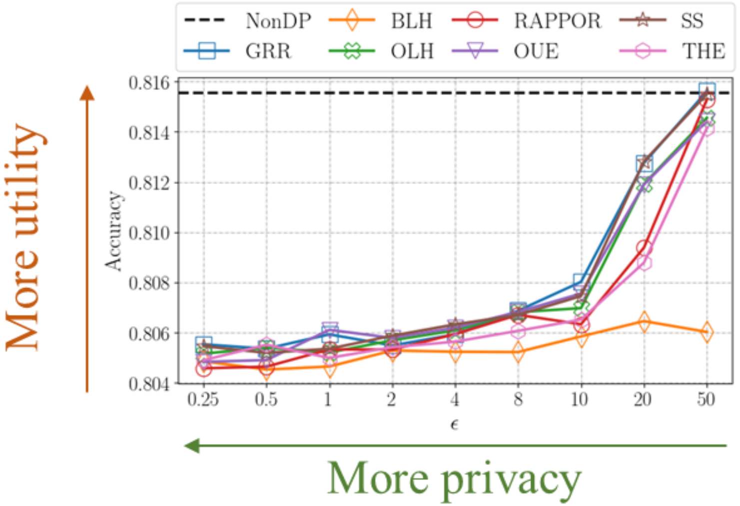
Dataset after LDP | ML algorithm | Output | Fairness metrics impacted?

Local DP (LDP) and Fairness: Friends! ~~or Foes~~?



Dataset after LDP   ML algorithm   Output   Fairness metrics impacted?





Differential Privacy (DP) and Fairness: Friends or Foes?

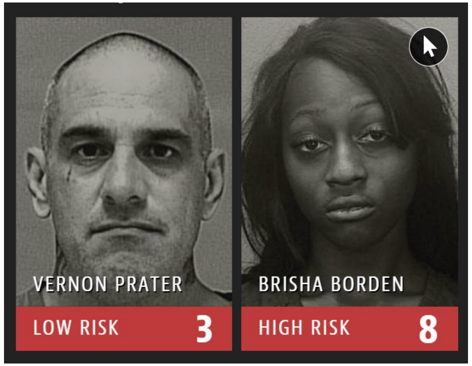
Paper	Task	Privacy	Details	Results
DP Has Disparate Impact on Model Accuracy (NeurIPS 2019)	Classification	Central DP	DP-SGD w/ same hyperparameters as the non-private baseline.	Foes
Robin Hood and Matthew Effects: DP Has Disparate Impact on Synthetic Data (ICML 2022)	Synthetic data generation + classification	Central DP	DP generative models w/ same hyperparameters as the non-private baseline.	Foes
An Empirical Analysis of Fairness Notions under DP (PPAI 2023)	Classification	Central DP	DP-SGD: search for optimal hyperparameters.	Minor impact
DP has Bounded Impact on Fairness in Classification (ICML 2023)	Classification	Central DP	DP-SGD: Theory.	Bounded impact
Fair Learning with Private Demographic Data (ICML 2020)	Classification	Local DP	LDP on single attribute + fairness mitigation mechanism.	
On the application and impact of ϵ-DP and fairness in ambulance engagement time prediction (ICLR 2023)	Classification	Local DP	LDP on multiple attributes.	Friends
Our (DBSec 2023)	Classification	Local DP	LDP on multiple attributes.	Friends

Outline

1. Motivation
2. **Background**
3. Problem Statement & Methods
4. Experimental Results
5. Conclusion & Perspectives

Fairness Metrics

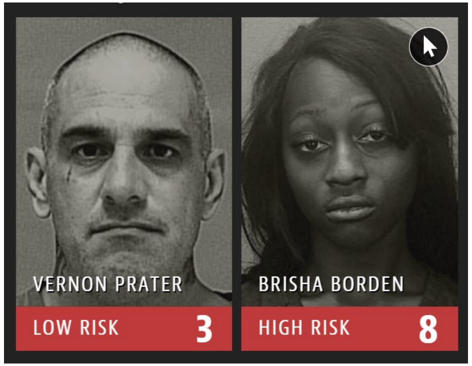
Fairness [Cambridge Dictionary]: The quality of treating people equally or in a way that is right or reasonable.



Fairness Metrics

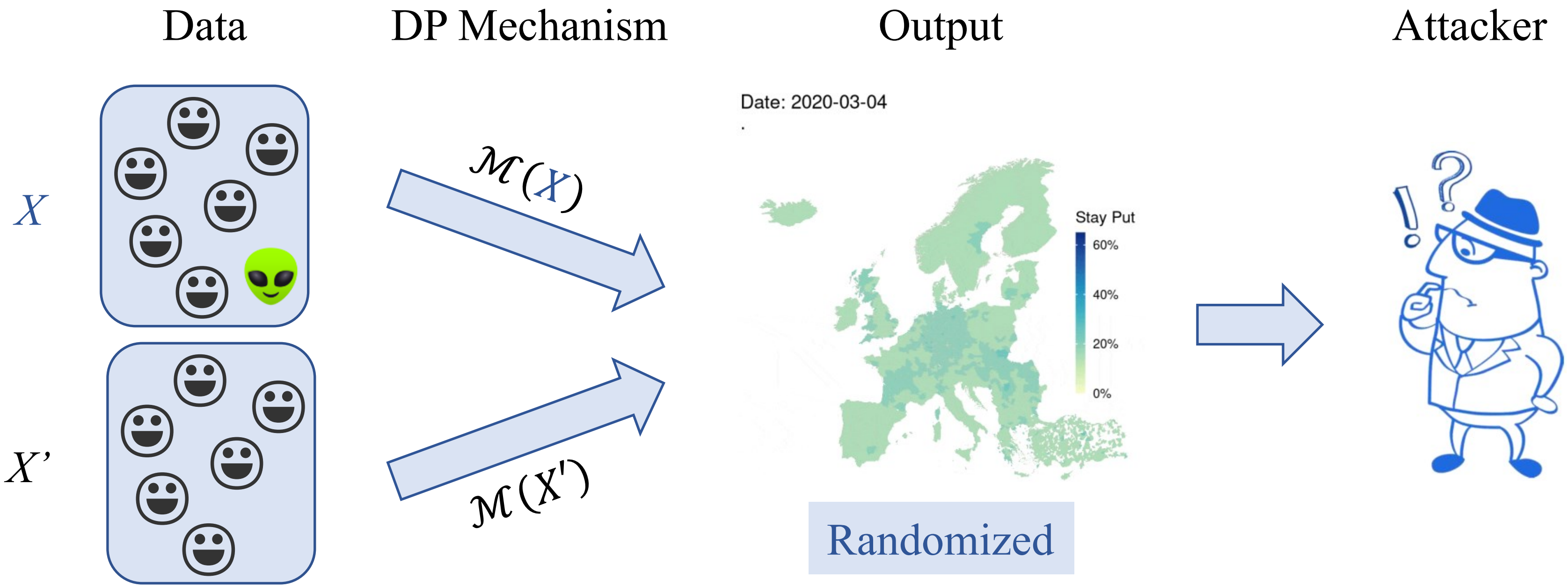
Fairness [Cambridge Dictionary]: The quality of treating people equally or in a way that is right or reasonable.

Protected attribute: $A_p \in \{0,1\}$
Target, predictor: $Y, \hat{Y} \in \{0,1\}$



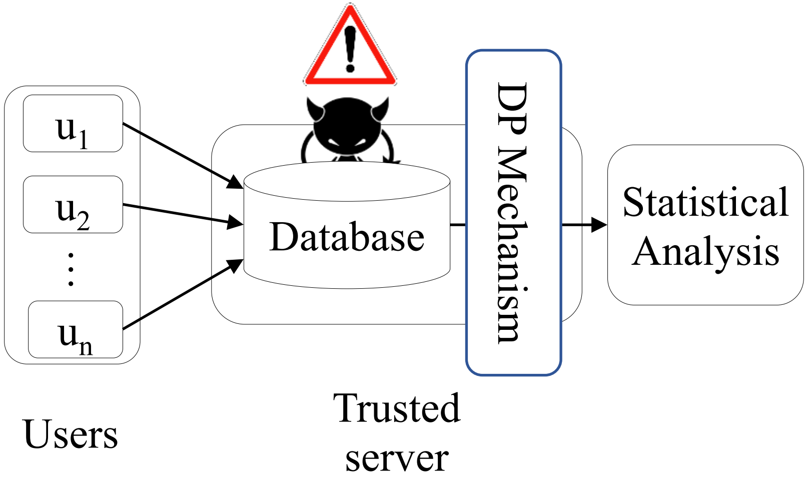
Fairness Metric	Equation	When Satisfied?
Disparate Impact (DI)	$\frac{\Pr[\hat{Y}=1 \mid A_p=0]}{\Pr[\hat{Y}=1 \mid A_p=1]}$	1
Statistical Parity Difference (SPD)	$\Pr[\hat{Y}=1 \mid A_p=1] - \Pr[\hat{Y}=1 \mid A_p=0]$	0
Equal Opportunity Difference (EOD)	$\Pr[\hat{Y}=1 \mid Y=1, A_p=1] - \Pr[\hat{Y}=1 \mid Y=1, A_p=0]$	0
Overall Accuracy Difference (OAD)	$\Pr[\hat{Y}=Y \mid A_p=1] - \Pr[\hat{Y}=Y \mid A_p=0]$	0

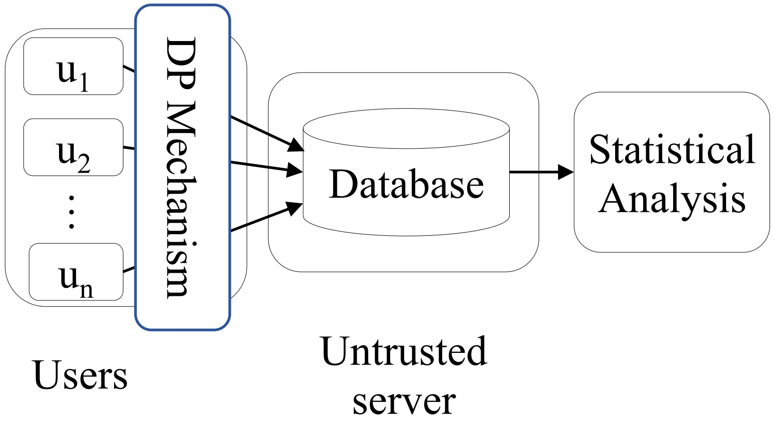
Differential Privacy (DP) [Dwork et al, 2006]



The attacker cannot tell if 👽 is in the sample

Differential Privacy (DP) [Dwork et al, 2006; Duchi et al, 2013]





Centralized DP:

✔ High utility.

✘ Need to trust the server.

✘✘ **Data breaches, data misuse, etc.**

Local DP (LDP):

✔ No need to trust the server.

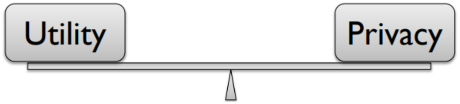
✘ Low utility.

LDP: Formal Definition & Properties [Duchi et al, 2013]

Def (ϵ-LDP). A randomized mechanism $\mathcal{M}$ satisfies ϵ-LDP, where $\epsilon \geq 0$, if for **any two inputs** $v, v' \in \text{Domain}(\mathcal{M})$ and for **any output** $z \in \text{Range}(\mathcal{M})$:

Privacy Loss

$$\frac{\Pr[\mathcal{M}(v) = z]}{\Pr[\mathcal{M}(v') = z]} \leq e^{\epsilon}$$

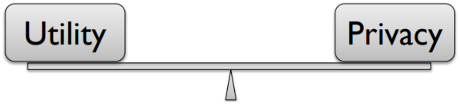


LDP: Formal Definition & Properties [Duchi et al, 2013]

Def (ϵ-LDP). A randomized mechanism $\mathcal{M}$ satisfies ϵ-LDP, where $\epsilon \geq 0$, if for **any two inputs** $v, v' \in \mathrm{Domain}(\mathcal{M})$ and for **any output** $z \in \mathrm{Range}(\mathcal{M})$:

$$\frac{\Pr[\mathcal{M}(v) = z]}{\Pr[\mathcal{M}(v') = z]} \leq e^{\epsilon}$$

Privacy Loss



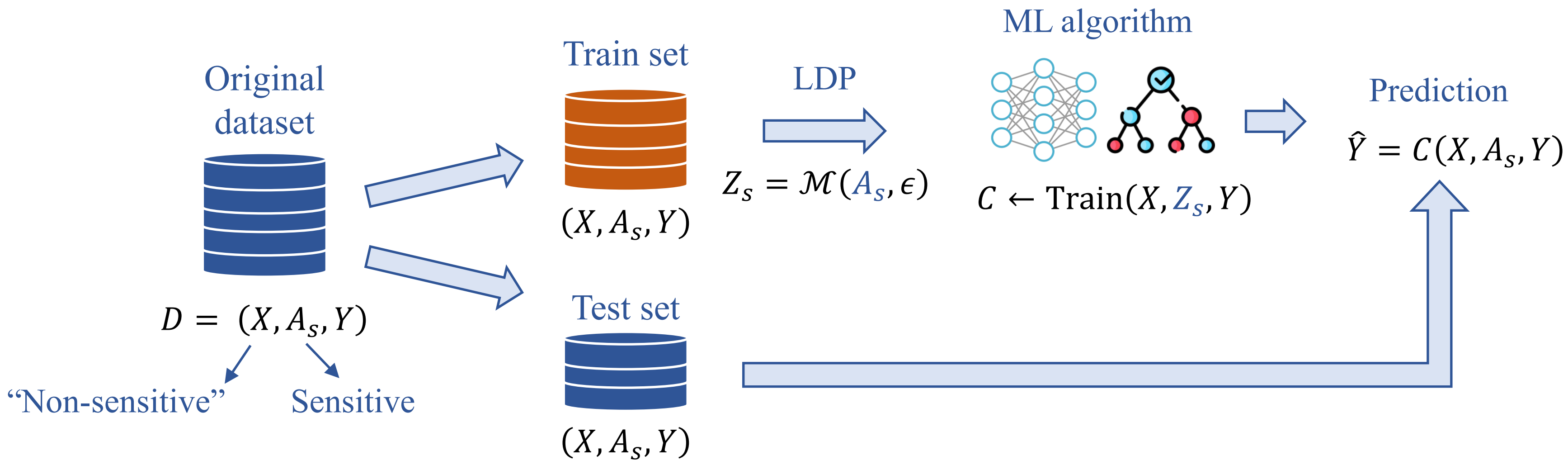
Fundamental (L)DP properties [Dwork et al, 2006]:

- **Post-processing** → if $\mathcal{M}$ is ϵ-LDP, then the composition $f(\mathcal{M})$ is ϵ-LDP for any f.
- **Composition** → Let $\mathcal{M}_1$ be a ϵ_1-LDP mechanism and $\mathcal{M}_2$ a ϵ_2-LDP mechanism. Then, the composed mechanism $\mathcal{M} = \big(\mathcal{M}_1(v), \mathcal{M}_2(v)\big)$ is $(\epsilon_1 + \epsilon_2)$-LDP.

Outline

1. Motivation
2. Background
3. **Problem Statement & Methods**
4. Experimental Results
5. Conclusion & Perspectives

Problem Statement



User's goal:

- Sanitize multiple sensitive attributes ($|A_s| \geq 2$) independently with ϵ-LDP.

Server's goal:

- Train a Machine Learning (ML) classifier on sanitized data (X, Z_s, Y).

Research Questions (RQs) & Assumptions

- RQ1: How does LDP pre-processing impacts fairness & utility?
- RQ2: How to better split the privacy budget ϵ for $d_s = |A_s|$ sensitive attributes?
- RQ3: Which LDP protocol lead to the best privacy-utility-fairness trade-off?

Research Questions (RQs) & Assumptions

- RQ1: How does LDP pre-processing impacts fairness & utility?
  - (Fairness) protected attribute A_p is always a sensitive attribute $A_p \in A_s$;
  - Empirical results w/ 3 datasets, 4 fairness metrics, and 4 utility metrics.
- RQ2: How to better split the privacy budget ϵ for $d_s = |A_s|$ sensitive attributes?
- RQ3: Which LDP protocol lead to the best privacy-utility-fairness trade-off?

Research Questions (RQs) & Assumptions

- RQ1: How does LDP pre-processing impacts fairness & utility?
  - (Fairness) protected attribute A_p is always a sensitive attribute $A_p \in A_s$;
  - Empirical results w/ 3 datasets, 4 fairness metrics, and 4 utility metrics.
- RQ2: How to better split the privacy budget ϵ for $d_s = |A_s|$ sensitive attributes?
  - State-of-the-art: Uniform splitting $\rightarrow \epsilon_j = \frac{\epsilon}{d_s}$ for $j \in A_s$;
  - Our solution: k-based $\rightarrow \epsilon_j = \frac{\epsilon \cdot k_j}{\sum_{i=1}^{d_s} k_i}$ for $j \in A_s$, $k_j = |A_j|$.

  ϵ-LDP following the sequential composition
- RQ3: Which LDP protocol lead to the best privacy-utility-fairness trade-off?

Research Questions (RQs) & Assumptions

- RQ1: How does LDP pre-processing impacts fairness & utility?
  - (Fairness) protected attribute A_p is always a sensitive attribute $A_p \in A_s$;
  - Empirical results w/ 3 datasets, 4 fairness metrics, and 4 utility metrics.
- RQ2: How to better split the privacy budget ϵ for $d_s = |A_s|$ sensitive attributes?
  - State-of-the-art: Uniform splitting $\rightarrow \epsilon_j = \frac{\epsilon}{d_s}$ for $j \in A_s$;
  - Our solution: k-based $\rightarrow \epsilon_j = \frac{\epsilon \cdot k_j}{\sum_{i=1}^{d_s} k_i}$ for $j \in A_s$, $k_j = |A_j|$.

ϵ-LDP following the sequential composition

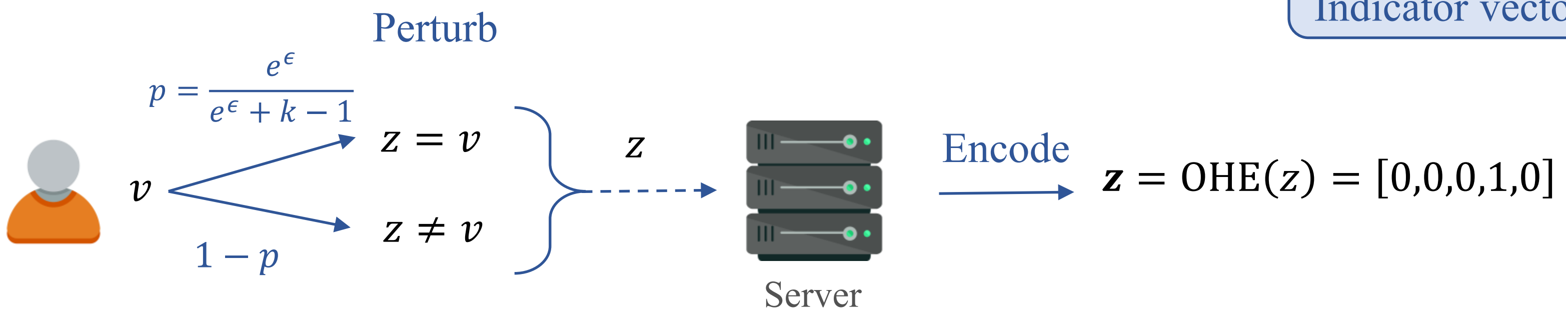
- RQ3: Which LDP protocol lead to the best privacy-utility-fairness trade-off?
  - Benchmarked 7 state-of-the-art LDP protocols;
  - Post-processed ϵ-LDP report for "homogeneous encoding" at the server side.

Inria | l'X ÉCOLE POLYTECHNIQUE

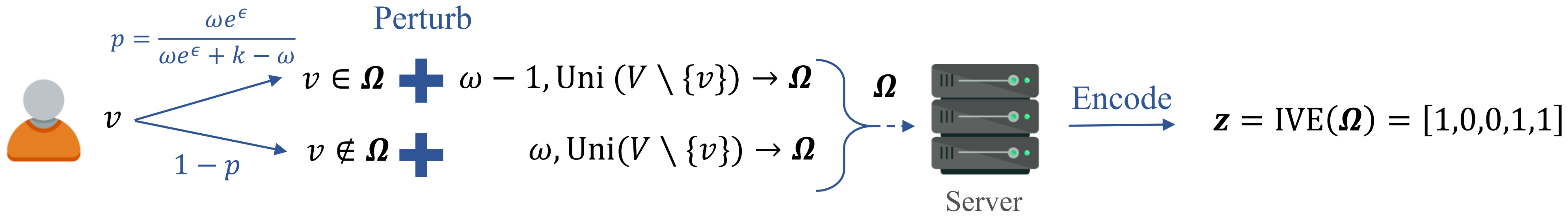
LDP Protocols & Server's "Homogeneous" Encoding

One-hot-encoding (OHE)
Indicator vector encoding (IVE)

Generalized Randomized Response (GRR)



Subset Selection (SS)



LDP Protocols & Server's "Homogeneous" Encoding

RAPPOR

$v \xrightarrow[\text{OHE}(v)]{\text{Encode}} \boldsymbol{v} = [0,0,0,1,0] \xrightarrow{\text{Perturb}} \boldsymbol{z} = [1,0,1,0,1] \xrightarrow{\boldsymbol{z}}$ Server

$$\Pr[\boldsymbol{z}_i = 1] = \begin{cases} \dfrac{e^{\epsilon/2}}{e^{\epsilon/2}+1} & \text{if } \boldsymbol{v}_i = 1, \\ \dfrac{1}{e^{\epsilon/2}+1} & \text{if } \boldsymbol{v}_i = 0. \end{cases}$$

Optimized Unary Encoding (OUE)

$v \xrightarrow[\text{OHE}(v)]{\text{Encode}} \boldsymbol{v} = [0,0,0,1,0] \xrightarrow{\text{Perturb}} \boldsymbol{z} = [1,0,0,1,1] \xrightarrow{\boldsymbol{z}}$ Server

$$\Pr[\boldsymbol{z}_i = 1] = \begin{cases} \dfrac{1}{2} & \text{if } \boldsymbol{v}_i = 1, \\ \dfrac{1}{e^{\epsilon}+1} & \text{if } \boldsymbol{v}_i = 0. \end{cases}$$

LDP Protocols & Server's "Homogeneous" Encoding

Local Hashing (LH)

Binary LH: $g = 2$
Optimal LH: $g = e^{\epsilon} + 1$

v Hash $H(v)$ → DFCA 54B4 BBEA 788A → mod g → $x \in [g]$

Perturb

$p = \frac{e^{\epsilon}}{e^{\epsilon} + g - 1}$: $z = x$

$1 - p$: $z \neq x$

$\langle \mathrm{H}, z \rangle$ → Server

Encode

$S(\langle H, z \rangle) = \{v \mid H(v) = z\}$

$\boldsymbol{z} = \mathrm{IVE}\big(S(\langle H, z \rangle)\big) = [1,0,0,0,1]$

Thresholding w/ Histogram Encoding (THE)

v Encode $\mathrm{OHE}(v)$ → $\boldsymbol{v} = [0,0,0,1,0]$ $\xrightarrow{\text{Perturb}}$ $\boldsymbol{z} = [1.3, \dots, -0.2]$ $\xrightarrow{\boldsymbol{z}}$ Server

$\boldsymbol{z}_i = \boldsymbol{v}_i + \mathrm{Lap}\left(\frac{2}{\epsilon}\right)$

Encode

$S(\boldsymbol{z}) = \{v \mid \boldsymbol{z}_v > \theta\}$

$\boldsymbol{z} = \mathrm{IVE}(S(\boldsymbol{z})) = [1,0,1,10]$

Outline

1. Motivation
2. Background
3. Problem Statement & Methods
4. **Experimental Results**
5. Conclusion & Perspectives

Setting of Experiments

Three datasets:

- Adult, ACSCoverage, LSAC.

Four fairness metrics:

- DI, SPD, EOD, AOD.

ML Classifier:

- LGBM w/ fixed hyperparameters;
- Train/test split as 80/20.

Seven LDP protocols:

- GRR, SS, RAPPOR, OUE, BLH, OLH, THE.

Two privacy budget splitting solutions:

- Uniform and k-based.

Stability: average over 20 runs

Fixed $|A_s| = 4$

Dataset	n	A_p	A_s, *domain size* k	Y
Adult	45849	gender	- gender, $k=2$ - race, $k=5$ - native country, $k=41$ - age, $k=74$	income
ACSCoverage	98739	DIS	- DIS, $k=2$ - AGEP, $k=50$ - SEX, $k=2$ - SCHL, $k=24$	PUBCOV
LSAC	20427	race	- race, $k=2$ - gender, $k=2$ - family income, $k=5$ - full time, $k=2$	pass bar

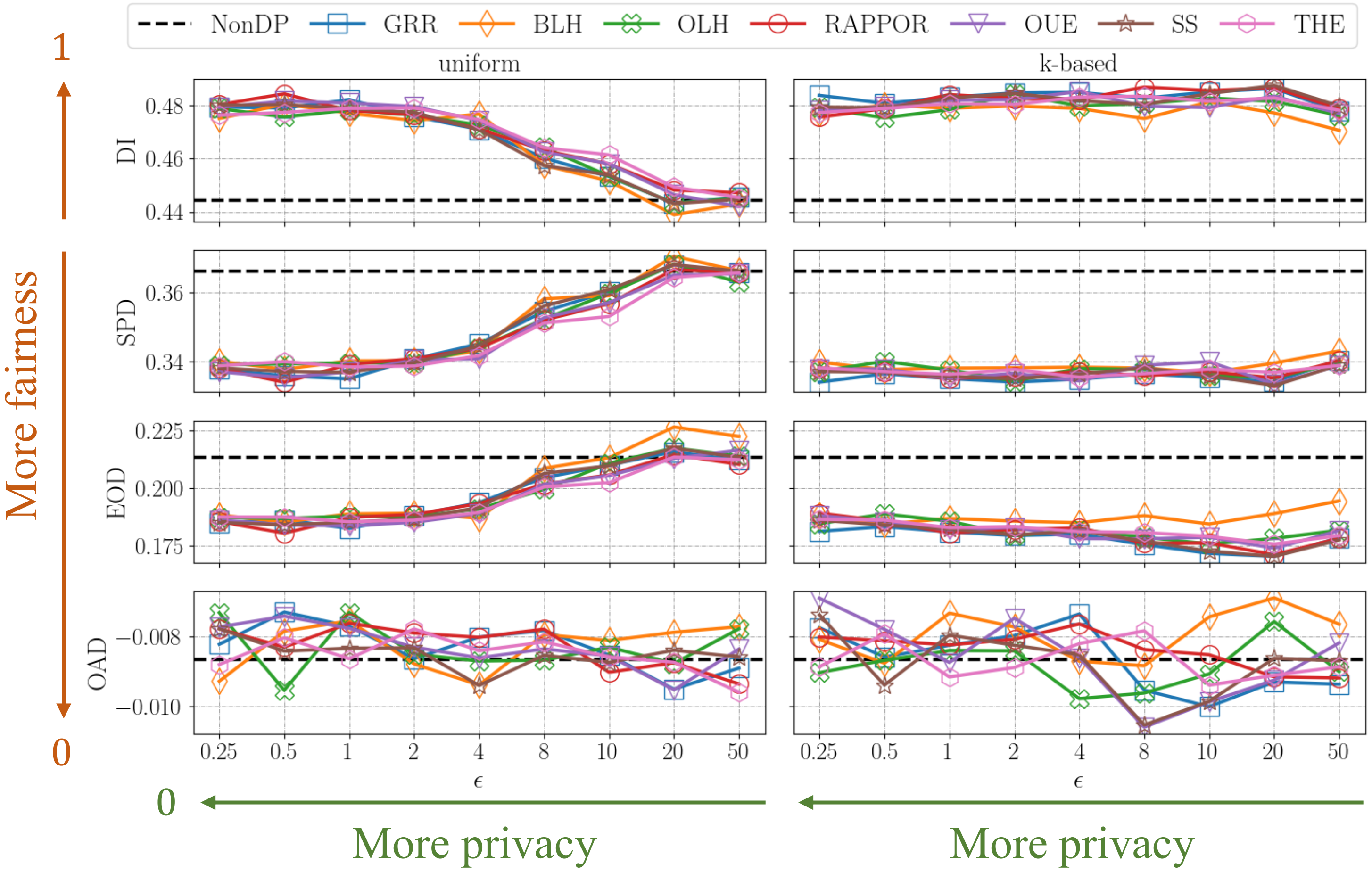
Impact of LDP on Fairness

$$\mathrm{DI} = \frac{\Pr[\hat{Y} = 1 | A_p = 0]}{\Pr[\hat{Y} = 1 | A_p = 1]} \rightarrow 1$$

$$\mathrm{SPD} = \Pr[\hat{Y} = 1 | A_p = 1] - \Pr[\hat{Y} = 1 | A_p = 0] \rightarrow 0$$

$$\mathrm{EOD} = \Pr[\hat{Y} = 1 | Y = 1, A_p = 1] - \Pr[\hat{Y} = 1 | Y = 1, A_p = 0] \rightarrow 0$$

$$\mathrm{OAD} = \Pr[\hat{Y} = Y | A_p = 1] - \Pr[\hat{Y} = Y | A_p = 0] \rightarrow 0$$



Impact of LDP on Fairness

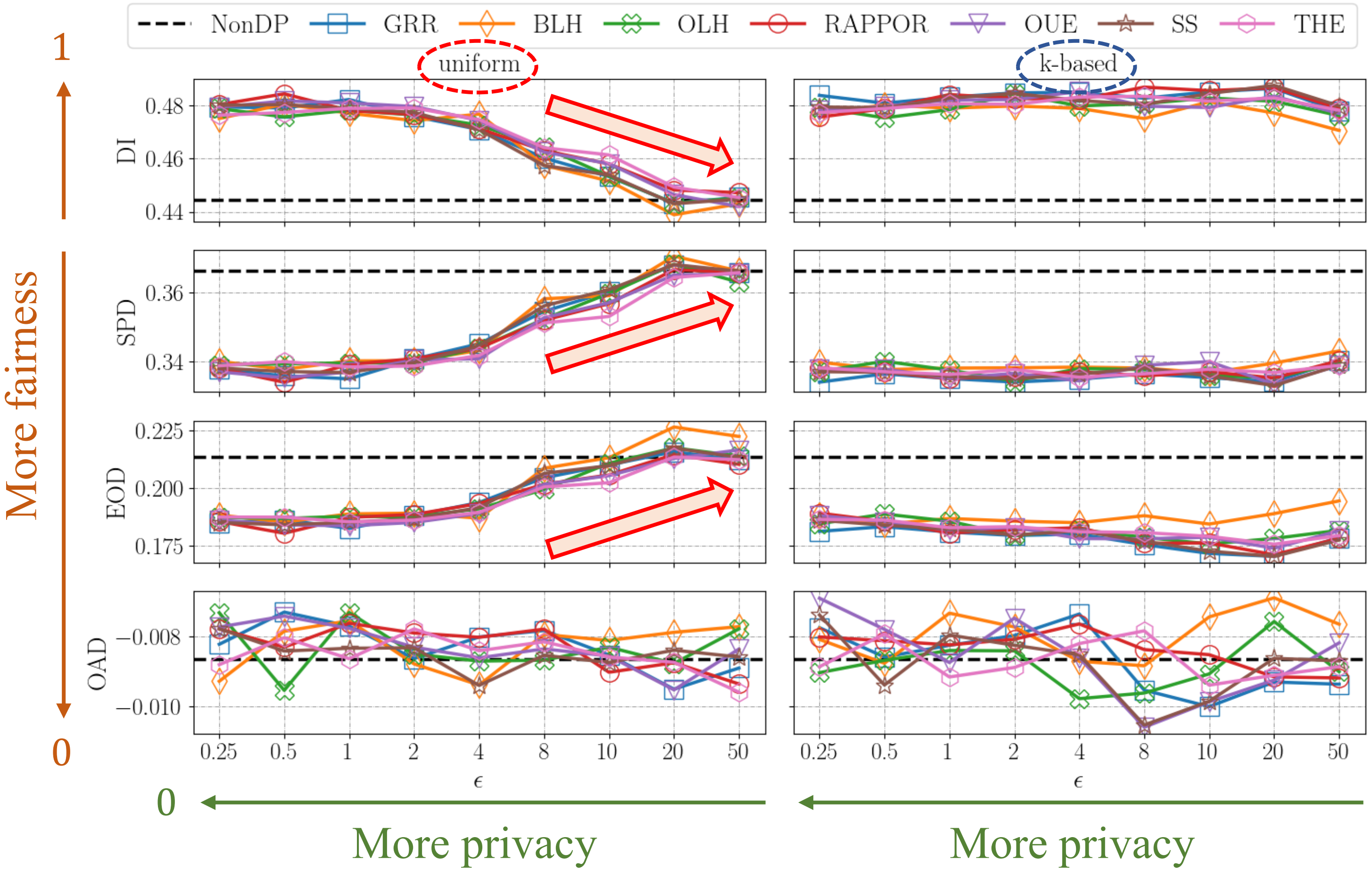
Uniform: goes towards the 'bad' baseline fairness metrics

$$\mathrm{DI} = \frac{\Pr[\hat{Y} = 1 | A_p = 0]}{\Pr[\hat{Y} = 1 | A_p = 1]} \rightarrow 1$$

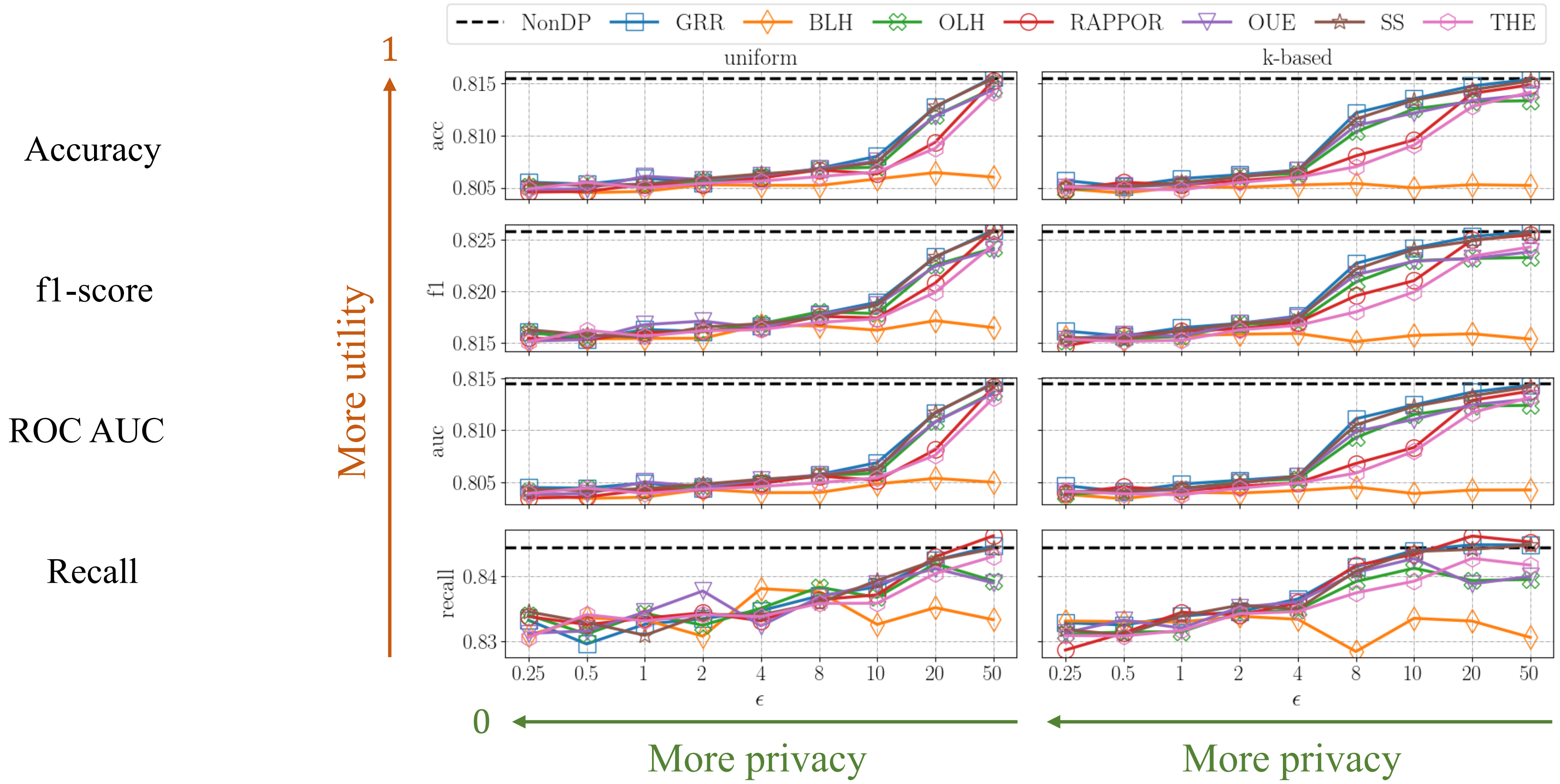
$$\mathrm{SPD} = \Pr[\hat{Y} = 1 | A_p = 1] - \Pr[\hat{Y} = 1 | A_p = 0] \rightarrow 0$$

$$\mathrm{EOD} = \Pr[\hat{Y} = 1 | Y = 1, A_p = 1] - \Pr[\hat{Y} = 1 | Y = 1, A_p = 0] \rightarrow 0$$

$$\mathrm{OAD} = \Pr[\hat{Y} = Y | A_p = 1] - \Pr[\hat{Y} = Y | A_p = 0] \rightarrow 0$$

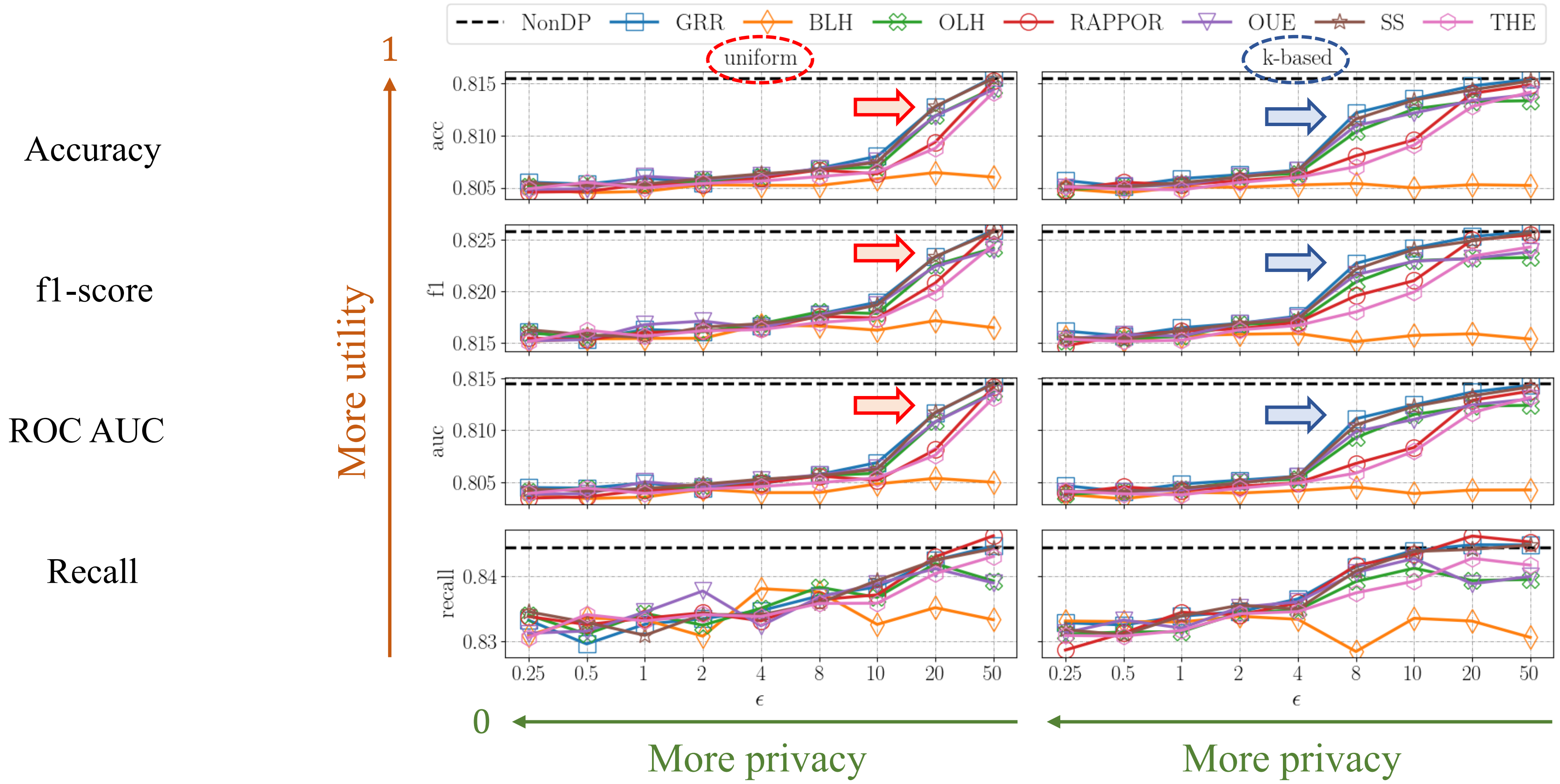


Impact of LDP on Utility



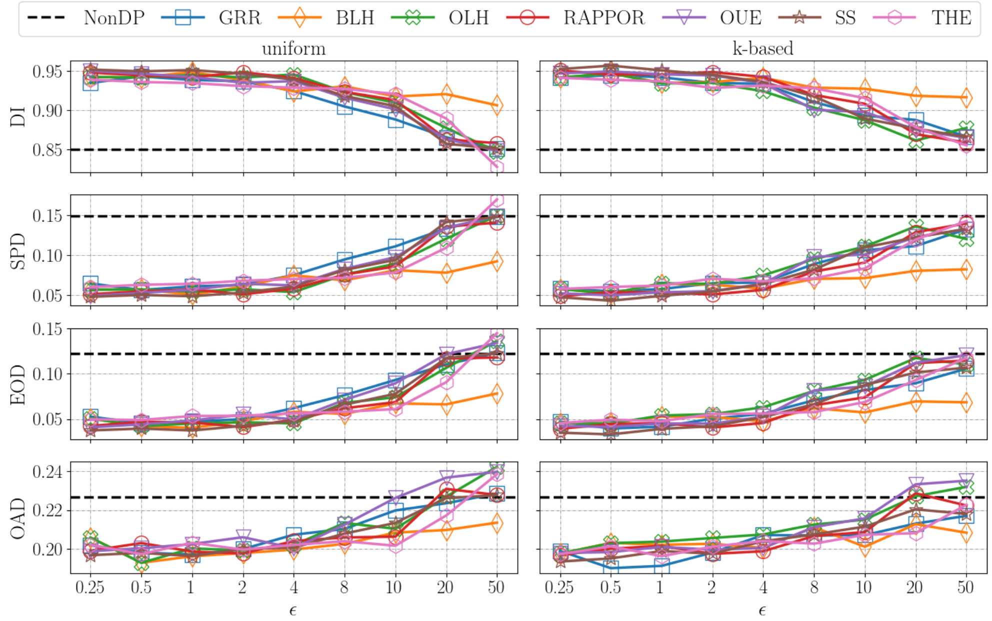
Impact of LDP on Utility

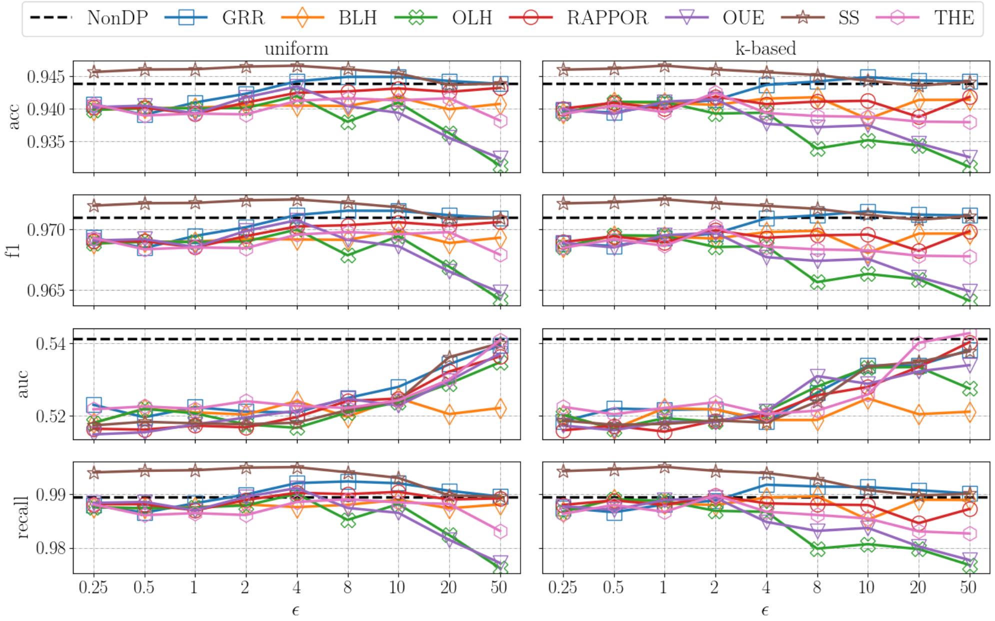
k-based: approaches faster the 'good' baseline utility metrics



Impact of LDP on Fairness & Utility: Generic? → Yes!

Appendix Experiments: $|A_s| = \text{Uniform}([2, 6])$.





More privacy

More privacy

Outline

Takeaway Messages

Conclusions:

- DP does not necessarely lead to worsened fairness in ML;
- (L)DP pre-processing positively affects fairness w/ minor utility impact;
- Our k-based solution leads to better privacy-utility-fairness trade-off;
- Mechanism w/ best privacy-utility-fairness trade-off: GRR and SS.

Takeaway Messages

Conclusions:

- DP does not necessarely lead to worsened fairness in ML;
- (L)DP pre-processing positively affects fairness w/ minor utility impact;
- Our k-based solution leads to better privacy-utility-fairness trade-off;
- Mechanism w/ best privacy-utility-fairness trade-off: GRR and SS.

Perspectives:

- Formalize our findings (*i.e.*, LDP & fairness trade-off);
- Introduce optimal mechanisms for privacy-fairness-aware ML;
- Study the impact of LDP pre-processing on different ML algorithms.

(Local) Differential Privacy has NO Disparate Impact on Fairness

Héber H. Arcolezi, Karima Makhlouf, and Catuscia Palamidessi

Inria and École Polytechnique (IPP), Palaiseau, France

PAPER

ARTIFACT

CONTACT

hharcolezi.github.io heber.hwang-arcolezi@inria.fr @hharcolezi