Label Differential Privacy can Release Unexpected Leakage

Hinata Sekiguchi†*, Shun Takagi‡, Satoshi Hasegawa‡, Marin Matsumoto†, and Masato Oguchi†
†Ochanomizu University, ‡LY Corporation
{hinata, marin}@ogl.is.ocha.ac.jp, oguchi@is.ocha.ac.jp
{shutakag, satoshi.hasegawa}@lycorp.co.jp

Abstract

Label Differential Privacy (Label DP) aims to improve utility by adding noise only to labels but becomes vulnerable to label leakage when features strongly correlate with labels. We define *Unexpected Leakage* to quantify the label leakage from features and show that Label DP fails to control it. We revisit Label DP with conditional features, as introduced by prior work, using this metric and show that it is well-suited to prevent Unexpected Leakage. Experiments reveal that Label DP generally achieves higher utility than a mechanism that prevents Unexpected Leakage, but in cases of strong feature-label correlations, the mechanism can outperform Label DP in terms of the utility-privacy trade-off. Our findings emphasize the importance of privacy standards that consider feature-label correlations and provide insights into balancing privacy and utility in differential privacy mechanisms.

1 Introduction

Differential Privacy (DP) has been widely adopted in numerous applications to mitigate privacy risks. (5; 12; 3; 7) However, conventional DP mechanisms apply noise to all user-related information, often leading to excessive noise that degrades the utility of the output. To solve this problem in scenarios where only labels contain sensitive information and features do not, Label Differential Privacy (Label DP) (6) has been introduced as a relaxed variant of DP. Label DP requires noise to be added solely to the output labels, allowing output features to remain unchanged, thereby enabling a substantial reduction in noise.

While Label DP has the advantage of utility, recent studies have pointed out that Label DP only prevents information leakage directly from noisy labels and neglects potential privacy risks arising from the correlation between labels and feature data (2). From the output raw feature, there exists a possibility that the sensitive label could be inferred from them, raising new privacy concerns.

This paper focuses on the fact that although such risks have been identified (2), the extent of label information leakage from features has not been rigorously quantified in prior research.

Our Contributions. In this work, we analyze the binary label information leakage from output features in the local setting (9) and make the following contributions:

- We define a novel metric, *Unexpected Leakage*, to quantify the extent of label information leakage due to output features.
- We show that the recently proposed privacy definition, Label Differential Privacy with conditional features (Label DP-CF) (1), prevents Unexpected Leakage, ensuring that label information cannot be inferred from output features
- We show that Label DP-CF is well-suited for protecting Unexpected Leakage by analyzing the relationship of Label DP-CF to Label DP and DP using Unexpected Leakage. Specifically, we theoretically prove that Label DP-CF imposes a stricter privacy guarantee compared to Label DP to prevent Unexpected Leakage while serving as a relaxed version of DP to utilize the non-sensitive features.
- Our experiments demonstrate that stronger feature-label correlations lead to greater unexpected leakage. While Label DP generally provides better utility than Local DP at the same leakage level, Local DP can outperform it when the feature-label correlation is particularly strong.

*Corresponding author:

2 Related Work

Existing work analyzed Label DP from the perspective of label inference attacks using the concept of Expected Attack Utility (EAU) (14; 2; 11). Busa-Fekete et al. (2) and Wu et al. (14) identified that Label DP mechanisms fail to fully protect label information due to inference attacks leveraging output features. Specifically, they demonstrated that Label DP cannot establish an upper bound on EAU. Moreover, Wu et al. (14) proposed the concept of *advantage* with the combination of EAUs, which represents information gain in a specific setting, and showed that Label DP appropriately bounds this gain. Busa-Fekete et al. (11) showed that Label DP controls the *advantage* better than learning-from-aggregate-labels techniques (15) with respect to the utility-privacy trade-off. However, these works do not explore another information gain that Label DP does not protect, which causes unbounded EAU.

In contrast, Busa-Fekete et al. (1) proposed Label Differential Privacy with conditional features (Label DP-CF). They showed that Label DP-CF prevents any label inference attack. However, it lacks an intuitive and interpretable assessment of Label DP-CF.

To address these limitations, we define a novel metric called *Unexpected Leakage* with new combination of EAUs, which measures the label information gain that Label DP does not protect. Specifically, we define the expected degree of label leakage as the EAU under randomized response (RR), which serves as a baseline privacy mechanism. The actual degree of label leakage is quantified by the EAU of an arbitrary mechanism $\mathcal{M}$. The difference between these two values defines what we refer to as *Unexpected Leakage*. By combining theoretical analysis and empirical validation, our study offers a novel approach to evaluating Label DP-CF and provides deeper insights into the practical feasibility of label privacy guarantees.

3 Preliminaries

Here, we introduce the privacy definitions that are related to our work.

Definition 3.1 (ε-Local Differential Privacy (4))**.** Given $\varepsilon \in \mathbb{R}^+$, $\mathcal{M}$ satisfies ε-DP if, for any two features $\forall x, x' \in \mathcal{X}$ and labels $\forall y, y' \in \mathcal{Y}$, and for any subset of outputs $Z \subseteq \mathcal{Z}$, the following inequality holds:

$$\frac{\Pr[\mathcal{M}(x, y) \in Z]}{\Pr[\mathcal{M}(x', y') \in Z]} \leq e^{\varepsilon}.$$

Definition 3.2 (ε-Label Differential Privacy (6))**.** $\mathcal{M}$ satisfies ε-Label DP if, for any feature $x \in \mathcal{X}$, any two labels $y, y' \in \mathcal{Y}$, and any subset of outputs $Z \subseteq \mathcal{Z}$, it holds that

$$\frac{\Pr[\mathcal{M}(x, y) \in Z]}{\Pr[\mathcal{M}(x, y') \in Z]} \leq e^{\varepsilon}.$$

Definition 3.3 ($(\varepsilon, \mathcal{P})$-Label Differential Privacy with conditional features (Label DP-CF) (1))**.** Given $\varepsilon \in \mathbb{R}^+$ and a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, a randomized mechanism $\mathcal{M} : \mathcal{X} \times \mathcal{Y} \mapsto \mathcal{Z}$ satisfies $(\varepsilon, \mathcal{P})$-Label DP with conditional features if, for any two labels $\forall y, y' \in \mathcal{Y}$ and for any subset of outputs $Z \subseteq \mathcal{Z}$, the following inequality holds:

$$\frac{\Pr_{X \sim \mathcal{P}(\mathcal{X}|y)}[\mathcal{M}(X, y) \in Z]}{\Pr_{X' \sim \mathcal{P}(\mathcal{X}|y')}[\mathcal{M}(X', y') \in Z]} \leq e^{\varepsilon}.$$

Label DP-CF is the definition naively derived by applying the notion of Definition 4.3 from prior work (1) to pure DP instead of Rényi DP. Label DP-CF supposes that the features are not assumed to be sensitive, but also not assumed to be public.

The (Individual) Expected Attack Utility of adversary $\mathcal{A} : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ which is modeled by a randomized function and mechanism $\mathcal{M}$ that perturbs an example from a distribution $\mathcal{P}$ ($\mathcal{P}(\mathcal{Y}|x)$ with a fixed feature x) is defined as:

$$\mathrm{EAU}(\mathcal{A}, \mathcal{M}, \mathcal{P}) = \Pr_{(X,Y) \sim \mathcal{P}}[\mathcal{A}(X, \mathcal{M}(X, Y)) = Y]\,.$$

$$\mathrm{IEAU}(\mathcal{A}, \mathcal{M}, \mathcal{P}, x) = \Pr_{Y \sim \mathcal{P}(\mathcal{Y}|x)}[\mathcal{A}(x, \mathcal{M}(x, Y)) = Y]\,.$$

4 Unexpected Leakage

Label DP improves model accuracy by reducing the noise compared to DP, assuming that features are not sensitive while labels remain private. This approach enhances data utility by ensuring label anonymity while allowing access to raw features.

However, features can still lead to label leakage if they are highly correlated with labels. For instance, if a dataset includes blood pressure as a feature and hypertension status as a label, the label can be inferred directly from the feature, even when anonymized. This demonstrates that label leakage

occurs beyond the guarantees of ε-Label DP, as an attacker can exploit feature-label correlations to infer private information.

Existing Label DP frameworks do not fully address this issue. To clarify this leakage, we define *Unexpected Leakage* as the degree of label leakage arising from features and formally quantify it in this study.

Definition 4.1 (Unexpected Leakage)**.** Given a mechanism $\mathcal{M}$ and a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, Unexpected Leakage (*UnExpected attack Advantage*) is defined as:

$$\mathrm{UEAdv}(\mathcal{M}, \mathcal{P}) = \sup_{\mathcal{A}} \mathrm{EAU}(\mathcal{A}, \mathcal{M}, \mathcal{P}) - \sup_{\mathcal{A}} \mathrm{EAU}(\mathcal{A}, \mathcal{M}_{\mathbf{WRR}}, \mathcal{P}).$$

Here, $\mathcal{M}_{\mathbf{WRR}}$ represents a mechanism that outputs only labels applied Warner's Randomized Response (RR) (13).

Similarly, when fixing the data point x, the Individual Unexpected Leakage (*Individual UnExpected attack Advantage*) is defined as:

$$\mathrm{IUEAdv}(\mathcal{M}, \mathcal{P}, x) = \sup_{\mathcal{A}} \mathrm{IEAU}(\mathcal{A}, \mathcal{M}, \mathcal{P}, x) - \sup_{\mathcal{A}} \mathrm{IEAU}(\mathcal{A}, \mathcal{M}_{\mathbf{WRR}}, \mathcal{P}, x).$$

Since $\mathcal{M}_{\mathbf{WRR}}$ produces randomized outputs using only label information, it prevents label leakage through correlations with features. Additionally, for binary labels, $\mathcal{M}_{\mathbf{WRR}}$ has been shown to be the optimal mechanism for label inference in terms of accuracy (8). Therefore, $\mathrm{EAU}(\mathcal{A}, \mathcal{M}_{\mathbf{WRR}}, \mathcal{P})$ can be considered the expected accuracy of the worst-case attacker under ε-Label DP. In contrast, when ε-Label DP allows highly correlated features to be released, the worst-case attacker's accuracy is represented by $\mathrm{IEAU}(\mathcal{A}, \mathcal{M}, \mathcal{P}, x)$.

From Definition 4.1, unexpected leakage is defined as the difference between these two values, and if it is greater than zero, unexpected leakage occurs.

Proposition 4.1. For the RR mechanism $\mathcal{M}_{RR}$, which release the original features and randomized labels, the maximum unexpected leakage IUEAdv depends on ε, the prior probability $\mathbb{P}(Y = y)$, and the posterior probability $\mathbb{P}(Y = y \mid x)$:

$$\mathrm{IUEAdv}(\mathcal{M}_{RR}, \mathcal{P}, x) \leq e^{\varepsilon} \cdot \{\mathbb{P}(Y = y \mid x) - \mathbb{P}(Y = y)\}.$$

This leads to the following corollary:

Corollary 4.2. If $r(x, y) = \frac{\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}}{\frac{\mathbb{P}(1-y|x)}{\mathbb{P}(1-y)}} > 1$, then $\mathcal{M}_{RR}$ exhibits unexpected leakage. Moreover, unexpected leakage can increase monotonically with $r(x, y)$.

Thus, even if a mechanism satisfies ε-Label DP, strong feature-label correlations can result in insufficient label protection. Since $r(x, y)$ and ε are independent, ε in Label DP fails to reliably control label privacy.

5 Analysis of Label DP-CF

In this section, we revisit Label DP-CF (1) from the perspective of Unexpected Leakage. First, we demonstrate that Label DP-CF prevents Unexpected Leakage. Then, we compare Label DP-CF with DP and Label DP within the framework of Unexpected Leakage, offering a clearer interpretation of Label DP-CF. These results collectively establish that Label DP-CF is a well-suited privacy definition to prevent Unexpected Leakage.

Unexpected Leakage in $(\varepsilon, \mathcal{P})$-Label DP-CF: We now establish the following theorem regarding unexpected leakage under $(\varepsilon, \mathcal{P})$-Label DP-CF.

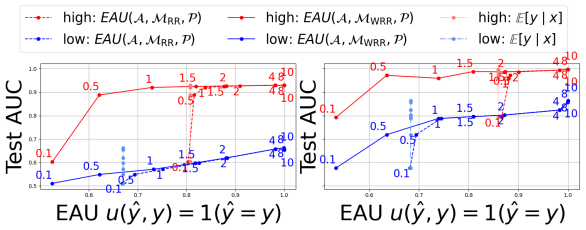
Theorem 5.1. Any mechanism satisfying $(\varepsilon, \mathcal{P})$-Label DP-CF does not exhibit unexpected leakage.

This result confirms that even in cases of strong feature-label correlations, label information remains protected under ε to the expected extent, equivalent to the protection provided by Warner's RR under ε.

Comparison of Label DP-CF with DP and Label DP: We compare $(\varepsilon, \mathcal{P})$-Label DP-CF against Label DP and DP, demonstrating that it provides stronger protection than Label DP while being less restrictive than DP with respect to Unexpected Leakage. This implies that it has more expressive power than ε-DP while still ensuring strict label protection under ε-Label DP.

Relation to ε-Label DP: The following theorem shows that, for datasets that do not exhibit unexpected leakage, $(\varepsilon, \mathcal{P})$-Label DP-CF is equivalent to Label DP.

Theorem 5.2. Any mechanism satisfying ε-Label DP-CF also satisfies ε-Label DP. The converse holds only if $r(x, y) = 1$.



(a) Breast cancer (b) Simulation dataset

Figure 1: Results on Unexpected Leakage.

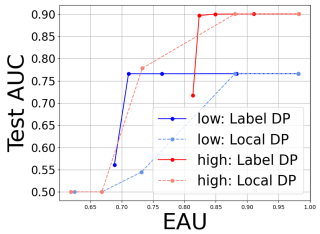


Figure 2: Label DP vs. Local DP. The values of ε for Label DP and DP are $\{8, 4, 2, 1, 0.1\}$ and $\{8, 4, 2, 1.4, 1\}$, respectively, from left to right.

This result establishes that $(\varepsilon, \mathcal{P})$-Label DP-CF serves as an improved version of ε-Label DP that accounts for unexpected leakage.

Additionally, the following corollary demonstrates that this concept relaxes the ε-DP definition.

Corollary 5.3. Any mechanism satisfying ε-DP also satisfies $(\varepsilon, \mathcal{P})$-Label DP-CF. The converse holds only if $r(x, y) = \infty$.

This suggests that mechanisms satisfying $(\varepsilon, \mathcal{P})$-Label DP-CF but not ε-DP can potentially achieve higher utility. Furthermore, this result indicates that ε-DP ensures label anonymity without unexpected leakage.

6 Experiment

Setting: Experiments on the simulated dataset followed the methodology of Wu et al. (14), while experiments on real-world data were conducted using the Breast Cancer Wisconsin (Diagnostic) Data Set (10).

6.1 Results on Unexpected Leakage

Figure 1 shows Unexpected Leakage on two datasets. The value next to each point denotes ϵ. The red line represents the results for the high dataset, where the relationship between labels and features is strong, while the blue line corresponds to the low dataset, where this relationship is relatively weaker. The dashed line indicates the leakage of a certain mechanism (EAU), the solid line represents the leakage level of Warner's RR, and the thin line shows the expected value of the conditional feature distribution $\mathbb{E}[y \mid x]$ for each dataset.

Figure 1b illustrates the results for the simulation dataset. The gap between the solid and dashed lines is larger for the red line, indicating that as the relationship between features and labels strengthens, the extent of unintended leakage increases.

Figure 1a presents the results for the real-world dataset. Since the exact probability $\mathbb{P}(Y = y \mid x)$ is unknown, we estimated it using an LLM(GPT-4o), and the corresponding results are shown by the dashed and thin lines. Similar to the simulation results, stronger feature-label relationships lead to greater unintended leakage.

6.2 Comparison of Label DP and Local DP

Figure 2 illustrates the overall leakage on the x-axis, which is the sum of unexpected and expected leakage, while the y-axis represents the test AUC, indicating utility. Overall, for the same level of leakage, Label DP achieves a higher test AUC, demonstrating its superiority as a method. However, in the lower right section of the red line, specifically when $\varepsilon_{\text{Label DP}} = 0.1$, Local DP achieves a higher test AUC at a comparable leakage level. This suggests that in datasets where the relationship between features and labels is strong, Local DP may sometimes outperform Label DP.

7 Conclusion

This study identified a fundamental limitation of conventional ε-Label Differential Privacy, demonstrating that it fails to fully prevent label leakage when labels are strongly correlated with publicly available features. To address this issue, we defined Unexpected Leakage to quantify the label leakage from features. Furthermore, we showed that the recently proposed privacy definition, Label DP-CF prevents Unexpected Leakage, ensuring that label information cannot be inferred from output features. Our experiments demonstrated that stronger feature-label correlations lead to greater unexpected leakage. While Label DP generally provided better utility than Local DP at the same leakage level, Local DP outperformed it when the feature-label correlation was particularly strong.

References

[1] Robert Busa-Fekete, Andres Munoz Medina, Umar Syed, and Sergei Vassilvitskii. 2023. Label differential privacy and private training data release. In *Proceedings of the Fortieth International Conference on Machine Learning (ICML 2023)*.

[2] Robert Istvan Busa-Fekete, Umar Syed, Sergei Vassilvitskii, et al. 2021. On the pitfalls of label differential privacy. In *NeurIPS 2021 Workshop LatinX in AI*.

[3] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. 2017. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580.

[4] Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming-Volume Part II*, pages 1–12. Springer-Verlag.

[5] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067.

[6] Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. 2021. Deep learning with label differential privacy. *Advances in Neural Information Processing Systems*, 34.

[7] Noah Johnson, Joseph P Near, and Dawn Song. 2018. Towards practical differential privacy for sql queries. *Proceedings of the VLDB Endowment*, 11(5):526–539.

[8] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2015. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR.

[9] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2011. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826.

[10] UCI Machine Learning Repository. 2016. Breast cancer wisconsin (diagnostic) data set. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. Accessed: 2025-01-04.

[11] Claudio Gentile Andres Munoz medina Adam Smith Robert Istvan Busa-Fekete, Travis Dick and Marika Swanberg. 2024. Auditing privacy mechanisms via label inference attacks. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[12] A. D. P. Team. 2017. Learning with privacy at scale.

[13] Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

[14] Ruihan Wu, Jin Peng Zhou, Kilian Q Weinberger, and Chuan Guo. 2023. Does label differential privacy prevent label inference attacks? In *International Conference on Artificial Intelligence and Statistics*.

[15] Felix X Yu, Dong Liu, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. 2013. svm for learning with label proportions. In *Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28*, pages III–504.

A Appendix

A.1 Proof of Proposition 4.1

Proposition (4.1)**.** The maximum value of IUEAdv depends on ε, $\mathbb{P}(y)$, and $\mathbb{P}(y \mid x)$.

Proof. IUEAdv is expressed as $\mathbb{P}(\mathbf{y} = y \mid \mathbf{X}, \tilde{\mathbf{y}}) - \mathbb{P}(\mathbf{y} = y \mid \tilde{\mathbf{y}})$. First, from Bayes' theorem, $\mathbb{P}(\mathbf{y} = y \mid \mathbf{X} = x, \tilde{\mathbf{y}} = \tilde{y})$ is given by

$$\frac{\mathbb{P}(\tilde{\mathbf{y}} = \tilde{y} \mid \mathbf{y} = y, x)\,\mathbb{P}(\mathbf{y} = y \mid x)}{\sum_{y' \in \{0,1\}} \mathbb{P}(\tilde{\mathbf{y}} = \tilde{y} \mid \mathbf{y} = y', x)\,\mathbb{P}(\mathbf{y} = y' \mid x)}$$

Since the output of $\mathcal{M}_{\mathbf{RR}}$ does not depend on x,

$$\mathbb{P}(\mathbf{y} = y \mid \mathbf{X} = x, \tilde{\mathbf{y}} = \tilde{y}) = \frac{\mathbb{P}(\tilde{\mathbf{y}} = \tilde{y} \mid \mathbf{y} = y)\,\mathbb{P}(\mathbf{y} = y \mid x)}{\sum_{y' \in \{0,1\}} \mathbb{P}(\tilde{\mathbf{y}} = \tilde{y} \mid \mathbf{y} = \mathbf{y} = y')\,\mathbb{P}(\mathbf{y} = y' \mid x)}$$

Furthermore, for simplicity, let $A := \mathbb{P}(\tilde{\mathbf{y}} = \tilde{y} \mid \mathbf{y} = y)$, $B := \mathbb{P}(\tilde{\mathbf{y}} = \tilde{y} \mid \mathbf{y} = 1 - y)$, and $p := \mathbb{P}(\mathbf{y} = y \mid x)$, then

$$\mathbb{P}(\mathbf{y} = y \mid \mathbf{X} = x, \tilde{\mathbf{y}} = \tilde{y}) = \frac{Ap}{Ap + B(1-p)}$$

Similarly, $\mathbb{P}(\mathbf{y} = y \mid \tilde{\mathbf{y}} = \tilde{y})$ is given by

$$\frac{\mathbb{P}(\tilde{\mathbf{y}} = \tilde{y} \mid \mathbf{y} = y)\,\mathbb{P}(\mathbf{y} = y)}{\sum_{y' \in \{0,1\}} \mathbb{P}(\tilde{\mathbf{y}} = \tilde{y} \mid \mathbf{y} = y')\,\mathbb{P}(\mathbf{y} = y')} = \frac{A\,\mathbb{P}(\mathbf{y} = y)}{A\,\mathbb{P}(\mathbf{y} = y) + B\,\mathbb{P}(\mathbf{y} = 1 - y)}$$

Letting $q = \mathbb{P}(\mathbf{y} = y)$,

$$\frac{Aq}{Aq + B(1-q)}.$$

Thus, IUEAdv is expressed as

$$\mathrm{IUEAdv} = \frac{Ap}{Ap + B(1-p)} - \frac{Aq}{Aq + B(1-q)}.$$

Considering $f(t) := \frac{At}{At+B(1-t)}$,

$$\text{IUEAdv} = f(p) - f(q)$$

By the mean value theorem,

$$f(p) - f(q) = \int_q^p f'(t)dt \leq \max_{t\in[0,1]} f'(t) \cdot (p-q)$$

Thus, we consider $\max_{t\in[0,1]} f'(t)$.

$$\begin{aligned} f'(t) &= \frac{A\{B + t(A-B)\} - At(A-B)}{\{B+(A-B)t\}^2} \\ &= \frac{AB}{\{B+(A-B)t\}^2}. \end{aligned}$$

Here, $f'(t)$ attains its minimum value B at $t = 0$ when $A > B$ and attains its minimum value A at $t = 1$ when $B > A$. Thus,

$$\max_{t\in[0,1]} f'(t) = \frac{AB}{\min(A,B)^2} = \max\left\{\frac{A}{B}, \frac{B}{A}\right\}$$

From the reversal probability of the randomized response mechanism, $\max\left\{\frac{A}{B}, \frac{B}{A}\right\} = e^\varepsilon$, so

$$f(p) - f(q) \leq e^\varepsilon (p-q)$$

which implies

$$\begin{aligned} \text{IUEAdv} &= \mathbb{P}(\mathbf{y} = y \mid \mathbf{X}, \tilde{\mathbf{y}}) - \mathbb{P}(\mathbf{y} = y \mid \tilde{\mathbf{y}}) \\ &\leq e^\varepsilon \cdot \{\mathbb{P}(\mathbf{y} = y \mid x) - \mathbb{P}(\mathbf{y} = y)\} \end{aligned}$$

□

A.2 Proof of Corollary 4.2

Corollary (4.2)**.** When $r(x,y) = \frac{\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}}{\frac{\mathbb{P}(1-y|x)}{\mathbb{P}(1-y)}} > 1$, $\mathcal{M}_{RR}$ exhibits unexpected leakage. Furthermore, unexpected leakage increases monotonically with respect to $r(x,y) > 1$.

Proof.

$$\frac{\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}}{\frac{\mathbb{P}(1-y|x)}{\mathbb{P}(1-y)}} > 1$$

Here, letting $p := \mathbb{P}(y \mid x), q := \mathbb{P}(y)$,

$$\begin{aligned} \frac{\frac{p}{q}}{\frac{1-p}{1-q}} &> 1 \\ \frac{p}{q} &> \frac{1-p}{1-q} \\ p - pq &> q - pq \\ p &> q \\ \therefore e^\varepsilon \cdot (p-q) &> 0 \end{aligned}$$

□

A.3 Proof of Theorem 5.2

Corollary (5.2)**.** A mechanism satisfying ε-Label DP-CF also satisfies ε-Label DP. The converse holds only if $r(x,y) = \frac{\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}}{\frac{\mathbb{P}(y'|x)}{\mathbb{P}(y')}} = 1$.

Proof. First, we prove that a mechanism $\mathcal{M}$ satisfying ε-Label DP-CF also satisfies ε-Label DP. For some mechanism $\mathcal{M}'$ satisfying ε-Label DP, there exists a (randomized) post-processing function f such that $f(\mathcal{M}'(X,y))$ is equivalent to $\mathcal{M}$. Thus, by the post-processing theorem of Label DP, $\mathcal{M}$ provides the same privacy guarantees as $\mathcal{M}'$. This implies that $\mathcal{M}$ satisfies ε-Label DP.

Next, we prove the converse.

$$\frac{\Pr[\mathcal{M}_{RR}(\mathbf{X},\mathbf{y}) = z \mid \mathbf{y} = y]}{\Pr[\mathcal{M}_{RR}(\mathbf{X},\mathbf{y}) = z \mid \mathbf{y} = y']} = \frac{\int_{x\sim\mathcal{P}(\mathcal{X}|y)} \Pr[\mathcal{M}_{RR}(\mathbf{X},\mathbf{y}) = z \mid (\mathbf{X},\mathbf{y}) = (x,y)]\,\mathbb{P}(x \mid y)dx}{\int_{x'\sim\mathcal{P}(\mathcal{X}|y')} \Pr[\mathcal{M}_{RR}(\mathbf{X},\mathbf{y}) = z \mid (\mathbf{X},\mathbf{y}) = (x',y')]\,\mathbb{P}(x' \mid y')dx'}$$

Since $\mathcal{M}_{RR}$ does not modify features, we have $x = x' = \tilde{x}$, leading to

$$\begin{aligned} &\frac{\Pr[\mathcal{M}_{RR}(\mathbf{X},\mathbf{y}) = z \mid \mathbf{y} = y]}{\Pr[\mathcal{M}_{RR}(\mathbf{X},\mathbf{y}) = z \mid \mathbf{y} = y']} \\ &= \frac{\Pr[\mathcal{M}_{RR}(\mathbf{X},\mathbf{y}) = (\tilde{x},\tilde{y}) \mid (\mathbf{X},\mathbf{y}) = (\tilde{x},y)]}{\Pr[\mathcal{M}_{RR}(\mathbf{X},\mathbf{y}) = (\tilde{x},\tilde{y}) \mid (\mathbf{X},\mathbf{y}) = (\tilde{x},y')]} \\ &\quad * \frac{\mathbb{P}(\tilde{x} \mid y)}{\mathbb{P}(\tilde{x} \mid y')} \\ &= \frac{\mathbb{P}(\tilde{y} \mid y)\,\mathbb{P}(\tilde{x} \mid y)}{\mathbb{P}(\tilde{y} \mid y')\,\mathbb{P}(\tilde{x} \mid y')} \\ &= \frac{\mathbb{P}(\tilde{y} \mid y)\frac{\mathbb{P}(y|\tilde{x})}{\mathbb{P}(y)}}{\mathbb{P}(\tilde{y} \mid y')\frac{\mathbb{P}(y'|\tilde{x})}{\mathbb{P}(y')}} \end{aligned}$$

Thus,

$$\begin{aligned} \frac{\frac{\mathbb{P}(y|\tilde{x})}{\mathbb{P}(y)}}{\frac{\mathbb{P}(y'|\tilde{x})}{\mathbb{P}(y')}} e^{-\varepsilon} &\leq \frac{\Pr[\mathcal{M}_{RR}(\mathbf{X},\mathbf{y}) = z \mid \mathbf{y} = y]}{\Pr[\mathcal{M}_{RR}(\mathbf{X},\mathbf{y}) = z \mid \mathbf{y} = y']} \\ &\leq \frac{\frac{\mathbb{P}(y|\tilde{x})}{\mathbb{P}(y)}}{\frac{\mathbb{P}(y'|\tilde{x})}{\mathbb{P}(y')}} e^{\varepsilon} \end{aligned}$$

Therefore, when $r(x,y) = \frac{\frac{\mathbb{P}(y|\tilde{x})}{\mathbb{P}(y)}}{\frac{\mathbb{P}(y'|\tilde{x})}{\mathbb{P}(y')}} \neq 1$, the mechanism does not satisfy ε-Label DP-CF, and as $r(x,y) > 1$ increases, the upper bound also increases.

□

A.4 Proof of Theorem 5.3

Corollary (5.3)**.** A mechanism $\mathcal{M}$ satisfying ε-DP also satisfies ε-Label DP-CF. The converse holds only if $r(x, y) = \infty$.

Proof. For any output subset $Z \in \mathcal{Z}$, we need to prove that if a mechanism $\mathcal{M}$ satisfies ε-Differential Privacy, then $\mathcal{M}$ also satisfies:

$$\begin{aligned}
&\Pr_{X\sim\mathcal{P}(\mathcal{X}|y)}[\mathcal{M}(X, y) \in Z] \leq \\
&e^{\varepsilon} \Pr_{X'\sim\mathcal{P}(\mathcal{X}|y')}[\mathcal{M}(X', y') \in Z] \\
&\Leftrightarrow \int_x \mathbb{P}(X = x \mid y) \Pr[\mathcal{M}(x, y) \in Z]dx \\
&\leq e^{\varepsilon} \int_{x'} \mathbb{P}(X = x' \mid y') \Pr[\mathcal{M}(x', y') \in Z]dx'
\end{aligned}$$

If a mechanism $\mathcal{M}$ satisfies ε-Differential Privacy, then for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$, we have:

$$\Pr[\mathcal{M}(x, y) \in Z] \leq e^{\varepsilon} \Pr[\mathcal{M}(x', y') \in Z]$$

Thus, we obtain:

$$\begin{aligned}
&\Pr[\mathcal{M}(x, y) \in Z] \\
&= \int_{x'} \mathbb{P}(X = x' \mid y') \Pr[\mathcal{M}(x, y) \in Z]dx' \\
&\leq e^{\varepsilon} \int_{x'} \mathbb{P}(X = x' \mid y') \Pr[\mathcal{M}(x', y') \in Z]dx' ,
\end{aligned}$$

Therefore, for each x,

$$\begin{aligned}
&\mathbb{P}(X = x \mid y) \Pr[\mathcal{M}(x, y) \in Z] \leq \mathbb{P}(X = x \mid y) \cdot \\
&\left(e^{\varepsilon} \int_{x'} \mathbb{P}(X = x' \mid y') \cdot \Pr[\mathcal{M}(x', y') \in Z]\, dx' \right),
\end{aligned}$$

Integrating over x,

$$\begin{aligned}
&\int_x \mathbb{P}(X = x \mid y) \Pr[\mathcal{M}(x, y) \in Z]dx \\
&\leq e^{\varepsilon} \left(\int_x \mathbb{P}(X = x \mid y)dx \right) \left(\int_{x'} \mathbb{P}(X = x' \mid y') \cdot \right. \\
&\left. \Pr[\mathcal{M}(x', y') \in Z]\, dx' \right),
\end{aligned}$$

Since $\int_x \mathbb{P}(X = x \mid y)dx = 1$, we obtain:

$$\begin{aligned}
&\int_x \mathbb{P}(X = x \mid y) \Pr[\mathcal{M}(x, y) \in Z]dx \\
&\leq e^{\varepsilon} \int_{x'} \mathbb{P}(X = x' \mid y') \cdot \Pr[\mathcal{M}(x', y') \in Z]\, dx'.
\end{aligned}$$

Therefore, if a mechanism $\mathcal{M}$ satisfies ε-DP, then it also satisfies ε-Label DP-CF.

Next, we prove the converse. Consider the following expression for mechanism $\mathcal{M}$:

$$\frac{\int_x \mathbb{P}(X = x \mid y) \Pr[\mathcal{M}(x, y) \in Z]dx}{\int_x \mathbb{P}(X = x \mid y') \Pr[\mathcal{M}(x, y') \in Z]dx}$$

If $r(x, y) = \infty$ and $\mathcal{M}$ does not satisfy ε-DP, this value becomes ∞. Thus, $\mathcal{M}$ must satisfy ε-DP. Conversely, if $r(x, y) < \infty$, then a mechanism $\mathcal{M}$ that does not satisfy ε-DP may still satisfy ε-Label DP-CF. Therefore, only when $r(x, y) = \infty$, a mechanism $\mathcal{M}$ satisfying ε-Label DP-CF also satisfies ε-DP.

□

A.5 Proof of Theorem 5.1

Theorem (5.1)**.** A mechanism satisfying ε-Label DP-CF does not exhibit unexpected leakage.

Proof. Let $\mathcal{M}$ be a mechanism that satisfies ε-Label DP-CF. Suppose that X and X' are random variables following the distributions $\mathcal{P}(\mathcal{X} \mid y)$ and $\mathcal{P}(\mathcal{X} \mid 1 - y)$, respectively. Then, the probability distributions $\mathcal{M}(X, y)$ and $\mathcal{M}(X', 1 - y)$ satisfy ε-indistinguishability. From this property, there exist probability distributions $Q(y)$ and $Q(1 - y)$ such that:

$$\mathcal{M}(X, y) = Q(y)\frac{\mathrm{e}^{\varepsilon}}{1 + \mathrm{e}^{\varepsilon}} - Q(1 - y)\frac{1}{1 + \mathrm{e}^{\varepsilon}}$$

$$\mathcal{M}(X', 1 - y) = Q(1 - y)\frac{\mathrm{e}^{\varepsilon}}{1 + \mathrm{e}^{\varepsilon}} - Q(y)\frac{1}{1 + \mathrm{e}^{\varepsilon}}$$

as shown in (8).

Thus, $\mathcal{M}$ can be interpreted as applying post-processing to the output $\tilde{y}$ of $\mathcal{M}_{RR}$ without observing x or y. This implies that the attacks feasible for $\mathcal{A}_{\text{informed}}$ are a subset of those feasible for $\mathcal{A}_{\text{naive}}$. Consequently, both IUEAdv and IUEAdv are at most 0, proving that unexpected leakage does not occur. □