The Fundamental Limits of Least-Privilege Learning

Theresa Stadler [1 *] **Bogdan Kulynych** [2 *] **Michael C. Gastpar** [1] **Nicolas Papernot** [3] **Carmela Troncoso** [1]

Abstract

The promise of least-privilege learning – to find feature representations that are useful for a learning task but prevent inference of any sensitive information unrelated to this task – is highly appealing. However, so far this concept has only been stated informally. It thus remains an open question whether and how we can achieve this goal. In this work, we provide the *first formalisation of the least-privilege principle for machine learning* and characterise its feasibility. We prove that there is a *fundamental trade-off* between a representation's utility for a given task and its leakage beyond the intended task: it is not possible to learn representations that have high utility for the intended task but, at the same time prevent inference of any attribute other than the task label itself. This trade-off holds *regardless* of the technique used to learn the feature mappings that produce these representations. We empirically validate this result for a wide range of learning techniques, model architectures, and datasets.

1. Introduction

The need to reveal data to untrusted service providers to obtain value from machine learning as a service (MLaaS) puts individuals at risk of data misuse and harmful inferences. The service provider observes raw data records at training or inference time and might abuse them for purposes other than the intended learning task. For instance, an image shared with a provider for the purpose of face verification might be misused to infer an individual's race and lead to discrimination (Citron & Solove, 2022).

Sharing Representations to Prevent Data Misuse. Collaborative learning and model partitioning claim to prevent such misuse in model training and inference MLaaS settings, respectively. In both cases, individuals share a feature representation of their raw data with the service provider; in the form of model updates in the collaborative learning setting (McMahan & Ramage, 2017; Hao, 2019) and of feature encodings in the model partitioning setting (Osia et al., 2018; Chi et al., 2018; Wang et al., 2018; Brown et al., 2022). Proponents of both techniques argue that, because individuals *only share a representation of their data*, and not the data itself, the service provider no longer has access to information that might be abused for purposes other than the intended task.

Unintended Feature Leakage. However, previous research shows that a passive adversary can misuse the shared representations to infer data attributes that are unrelated to the learning task, or even reconstruct data records (Boenisch et al., 2023; Ganju et al., 2018; Melis et al., 2019; Song & Shmatikov, 2019). For instance, Song & Shmatikov (2019) show that features extracted from a gender classification model also reveal an individual's race. Even higher-layer features, that are assumed to be more learning-task specific, might lead to such unexpected inferences (Mo et al., 2021). These examples show that limiting data access to feature representations does not necessarily prevent unintended information leakage and thus, does not fully mitigate the risk of data misuse associated with attributes other than the learning task.

Least-Privilege Learning. Some works (Osia et al., 2018; Melis et al., 2019; Brown et al., 2022) suggest that the solution to this issue is to ensure that the feature mappings that produce such representations follow the *least-privilege principle* (LPP). That is, to enforce that the representations shared with the service provider *only contain information relevant to the learning task, and nothing else*. Previous work repeatedly suggests least-privilege learning (LPL) as a promising avenue to prevent data misuse. Yet, it has only been described informally and lacks a precise definition.

Contributions. We make the following contributions:

I) We provide the first formalisation of the least-privilege principle for machine learning as a variant of the generalized Conditional Entropy Bottleneck problem (Fischer, 2020). Our formalisation enables us to charac-

[*]Equal contribution [1]EPFL, Lausanne, Switzerland [2]Lausanne University Hospital & University of Lausanne, Switzerland [3]University of Toronto & Vector Institute, Toronto, Canada. Correspondence to: Theresa Stadler <theresa.stadler@epfl.ch>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

terise the limits of unintended information leakage in MLaaS settings.

II) We observe that any feature representation must at least leak all information about the input data that can be inferred from the learning task label itself. We show experimentally that this *fundamental leakage* can reveal information that is not intuitively related to the intended task and that could be misused for harmful inferences.

III) We formally prove a fundamental trade-off: under realistic assumptions on the data distribution, it is not possible to learn feature representations that have high utility for the intended task but, at the same time, restrict information leakage about data attributes other than the intended task to its fundamental leakage.

IV) We experimentally demonstrate this trade-off across learning techniques, model architectures, and datasets. We show that as long as the representations have utility for their intended task, there exist attributes other than the task label that can be inferred from the shared representations and thus violate the least-privilege principle.

Related Work. Prior works refer to the LPP but so far only study the problem of *attribute obfuscation* for a *single, fixed sensitive attribute*. For instance, Song & Shmatikov (2019) experimentally observe that current censoring techniques can only prevent a model from learning a sensitive attribute at a cost in model utility. Their claim that "overlearning [of sensitive attributes] is intrinsic" (Song & Shmatikov, 2019) to machine learning models is derived from empirical observations on a single learning task and sensitive attribute for a small set of censoring techniques. Zhao et al. (2020) formally derive a lower bound on the trade-off between a model's performance on its intended learning task and hiding a *fixed sensitive attribute*.

While these results characterise the trade-off between a representation's utility and leakage with respect to a specific attribute, they do not actually evaluate whether it is possible to learn useful representations that fulfil the LPP. They focus on a single attribute, whereas the LPP demands that the representations shared with the service provider should prevent inference of *any information* other than the intended task (Melis et al., 2019; Brown et al., 2022). Furthermore, neither Song & Shmatikov (2019) nor Zhao et al. (2020) consider that useful representations must at least leak any sensitive information already revealed by the task label itself. As we argue in Section 3, such *fundamental leakage* is crucial to consider when formalising LPL.

2. Problem Setup

We consider the common MLaaS setting in which individual data owners, or users, share their data with a service provider for a target task, e.g., model training or inference. Users agree to the usage of their data for the intended task but want to prevent data misuse through the service provider. In this section, we formalize this problem setup.

Notation. Let $X, Y, S \sim P_{X,Y,S}$ be a set of random variables distributed according to $P_{X,Y,S}$ where $X \in \mathbb{X}$, and $Y \in \mathbb{Y}$ are, respectively, an *example* and its *learning task label*, and $S \in \mathbb{S}$ is a *sensitive attribute*. For any three random variables X, Y, W, we denote by $Y - X - W$ *a Markov chain*, which is equivalent to stating that: $Y \perp\!\!\!\perp W \mid X$.

Assumptions on the Data Distribution. To make our formal analyses tractable, we assume that the spaces $\mathbb{X}, \mathbb{Y}, \mathbb{S}$ are discrete and finite; and that the data domain is non-trivial ($|\mathbb{X}| > 1$ and $|\mathbb{Y}| > 1$) and has full support. We further make the following assumption about the data distribution:

Assumption A (Strictly positive posterior)**.** *We say that the* posterior distribution*, $P_{Y|X}$, is strictly positive if for any $x \in \mathbb{X}, y \in \mathbb{Y}$ we have $P_{Y|X}(y \mid x) > 0$.*

This assumption is realistic in settings where there exists inherent uncertainty about the ground truth label of a given example. Examples include the presence of *label noise* introduced by the labelling process (Song et al., 2022), and, under the Bayesian interpretation of probability, task labels that are subjective. This is the case, for instance, in many MLaaS applications, such as emotion recognition or prediction of face attributes (e.g., smiling) which come with uncertainty and labelling unreliability (Raji et al., 2021).

Problem Setup. In the MLaaS setting, individual *users* hold a set of *data records* (x_i, y_i) for $i = 1, \ldots, N$ and agree to share their data with a *service provider* for a specific *purpose*. For example, a user might be willing to reveal an image of their face to the provider of a biometric face recognition system to prove their identity. However, if the user directly reveals the original image X, the service provider can *misuse the shared data* for purposes other than the intended task, i.e., to verify the user's identity.

To prevent such data misuse, users share a *representation* of their data that restricts information leakage to the necessary minimum for the *intended purpose* (see Fig. 1). For instance, in the previous example, a representation that only reveals features relevant to recognising a user's identity. The goal hence is to find a possibly randomized *feature mapping* $f_E(\cdot)$ that maps input X to a representation $Z = f_E(X)$ that both has *high utility for the intended task* Y but *prevents inference of any other data attribute* S. We formalise this goal and characterise its feasibility in the next section.

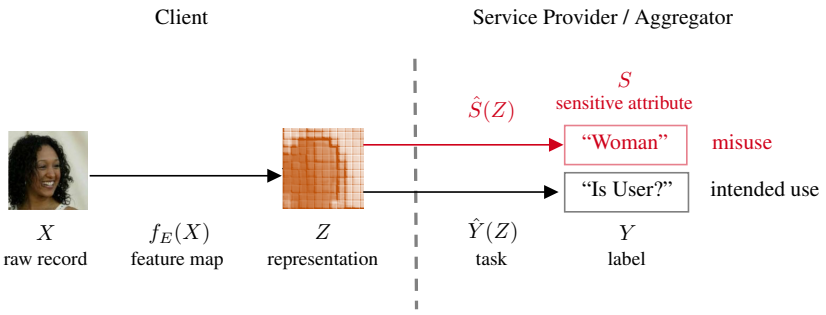

Figure 1: To prevent potential data misuse in a MLaaS setting, users share a representation of their data. These representations should be useful to achieve the intended purpose (verify users' identity) but prevent inference of other data attributes (users' gender) that might lead to harms, such as discrimination.

Training vs. Inference Settings. The example above describes a MLaaS *inference* setting in which users directly share a representation of their data for the purpose of predicting the intended task label. However, the same problem setup applies to the collaborative MLaaS *training* setting in which users share gradient updates instead of their raw data to prevent data misuse. As Melis et al. (2019) show, these "gradient updates can [...] be used to infer feature values" and thus are just a version of the model-generated representations $Z = f_E(X)$.

3. The Least-Privilege Principle in Machine Learning

The Least-Privilege Security Principle. The LPP is a design principle for building secure information systems introduced by Saltzer & Schroeder (1975). Its definition demands that "Every program and every user of the system should operate using the least set of privileges necessary to complete the job." In secure systems engineering, a privilege is a clearly defined action that an actor in the system is authorised to carry out. Transferring this concept to the machine learning domain is not trivial. To do so, we have to first quantise the learning process into a set of smaller privileges and then define what is the minimum set of privileges needed to carry out a learning task.

We approach this problem through the lens of *attribute inference*: The goal of the user is to allow inference of the task attribute Y from representation Z but prevent inference of any data attribute S other than the learning task. We define a *privilege as the ability to learn the value of a particular attribute*, and formalise the LPP in terms of inference gain about data attributes other than the learning task.

3.1. Utility Measure

The representation shared with the service provider should be useful for the intended purpose, i.e., contain features relevant to correctly infer task label Y. One common way to formalise this objective is the *mutual information* between the representations and the task label $I(Y;Z) = I(Y;f_E(X))$ (Alemi et al., 2016). The higher the information between Z and Y, the more useful the representation will be for the intended task. We make this notion more general by using a variant of mutual information known as Arimoto's α-information (Arimoto, 1977), which we denote as $I_\alpha(Y;Z)$. In particular, we consider two relevant instantiations of α: $\alpha = 1$ and $\alpha = \infty$. In the case that $\alpha = 1$, Arimoto's information is equal to classical Shannon's mutual information $I_\alpha(Y;Z) = I(Y;Z)$, described previously. In the case that $\alpha = \infty$, Arimoto's information is the *multiplicative gain* in accuracy of predicting the task label Y from the learned representations over baseline guessing (Liao et al., 2019):

$$I_\infty(Y;Z) \triangleq \log \frac{\Pr[Y = f_C(Z)]}{\Pr[Y = \hat{Y}]} \tag{1}$$

where $\log(\cdot)$ is the base-2 logarithm, $f_C(Z)$ denotes the prediction for task label Y of a classifier f_C based on representation Z, and $\hat{Y} \triangleq \arg\max_{y \in \mathbb{Y}} \Pr[Y = y]$ represents the baseline guess based on the most common attribute value. The latter notion of $I_\infty(Y;Z)$ is especially useful to evaluate utility of representations for classification tasks, as it is normalised prediction accuracy.

3.2. Leakage Measure

As described in Section 2, users want to restrict information leakage about data attributes other than the intended task. We evaluate such unintended leakage through the success of an *inference adversary* $g(W)$ that tries to infer the value of a sensitive attribute S from information W. In line with standard practices in security and privacy, we use Bayes-optimal adversaries (Sablayrolles et al., 2019; Chatzikokolakis et al., 2023) that achieve optimal inference accuracy and hence

measure worst-case inference risk:

$$\hat{S}(W) \triangleq \arg\max_{g:\, \mathbb{W}\to\mathbb{S}} \Pr[S = g(W)]. \tag{2}$$

We use $\hat{S}$ without the argument to denote the majority class baseline guess: $\hat{S} \triangleq \arg\max_{s\in\mathbb{S}} \Pr[S = s]$.

We measure unintended leakage about sensitive attribute S through an adversary's multiplicative gain (Arimoto's information with $\alpha = \infty$). Formally, the gain of an adversary $\hat{S}(W)$ who has access to information W over their baseline guess is defined as:

$$I_\infty(S; W) \triangleq \log \frac{\Pr[S = \hat{S}(W)]}{\Pr[S = \hat{S}]}. \tag{3}$$

We denote the gain of an adversary $\hat{S}(W, W')$ with access to two sources of information W and W' over their guess with only one source of information W' as:

$$I_\infty(S; W \mid W') \triangleq \log \frac{\Pr[S = \hat{S}(W, W')]}{\Pr[S = \hat{S}(W')]}. \tag{4}$$

We use these two measures (Eq. (1) as a representation's utility and Eq. (4) as its unintended leakage) to formalise the LPP in machine learning.

3.3. Strawman Approach: Unconditional Least-Privilege Principle

The promise of LPL, as suggested by prior work (Osia et al., 2018; Melis et al., 2019; Brown et al., 2022), is to find a feature representation that only contains information relevant to the "purpose for which it was designed and *nothing else*" (Brown et al., 2022). To formalise the LPP, in contrast to related works on attribute obfuscation, we hence cannot assume the sensitive attribute S to be fixed. Instead, we need to assume that *any attribute* — e.g., any function of the input — other than the learning task $S \neq Y$ is sensitive and its inference might lead to harm. Formally:

Definition 1 (Unconditional LPP)**.** *Given a data distribution $P_{X,Y}$, a feature map $f_E(X) = Z$ satisfies the unconditional LPP with parameter γ if for any attribute $S \neq Y$ which follows the Markov chain $S - X - Z$, the attribute inference gain is bounded:*

$$\frac{\Pr[S = \hat{S}(Z)]}{\Pr[S = \hat{S}]} \leq 2^\gamma \tag{5}$$

Equivalently:

$$\sup_{S\neq Y:\, S-X-Z} I_\infty(S; Z) \leq \gamma \tag{6}$$

Many previous works suggest that it is possible to find a feature map $f_E(X) = Z$ that fulfils the unconditional LPP, and at the same time produces representations with high utility for the learning task (Osia et al., 2018; Melis et al., 2019; Brown et al., 2022). To address the question whether it is possible to achieve high utility for the intended task and simultaneously satisfy the unconditional LPP, we formally characterise this trade-off:

Theorem 1 (Unconditional LPP and Utility Trade-Off)**.** Suppose that $P_{Y|X}$ is strictly positive (Assumption A). Then, for $\alpha \in \{1, \infty\}$, the following two properties cannot hold at the same time:

I) $Z = f_E(X)$ satisfies the unconditional LPP with parameter γ

II) $I_\alpha(Y; Z) > \gamma$

We provide a proof of this and all of the following formal statements in Appendix A. This result implies that whenever a representation has a certain utility for the learning task with $I_\alpha(Y, Z) > \gamma$, there exists a sensitive attribute for which an adversary's inference gain is at least as large with $I_\infty(S, Z) \geq \gamma$. In fact, under Definition 1, it is easy to construct this attribute to be infinitesimally close to the task label Y but not quite match it. In the next section, we provide an alternative formalisation of the LPP that captures the requirement $S \neq Y$ yet precludes these cases.

3.4. Formalisation of the Least-Privilege Principle

In the previous section, we show that the unconditional LPP, i.e., hiding *all* information about X, comes with a stringent trade-off. But this is an unnecessarily restrictive goal. To use a service, users must be willing to reveal to the service provider at least the intended result of the computation, i.e., the task label Y. As a consequence, they cannot conceal from the service provider *any information that can be inferred from Y itself*. This information hence defines the *least privilege* that can be given to the service provider, i.e., the minimum access to data attributes that must be granted to carry out a task. We call this information the *fundamental leakage* of the task. For a given attribute S, the fundamental leakage equals $\Pr[S = \hat{S}(Y)]$.

We propose a formalisation of the LPP that only demands that sharing a record's feature representation $Z = f_E(X)$ does not reveal *more information* about a sensitive attribute S than publishing Y itself:

Definition 2 (LPP)**.** *Given a data distribution $P_{X,Y}$, a feature map $f_E(X) = Z$ satisfies the LPP with parameter γ if for any attribute S which follows the Markov chain $S - (X, Y) - Z$, the attribute inference gain from observing (Z, Y) over the fundamental leakage is bounded:*

$$\frac{\Pr[S = \hat{S}(Z, Y)]}{\Pr[S = \hat{S}(Y)]} \leq 2^\gamma \tag{7}$$

Equivalently:

$$\sup_{S:\ S-(X,Y)-Z} I_\infty(S; Z \mid Y) \leq \gamma \tag{8}$$

Notably, the quantity constrained by the LPP is known as *maximal leakage* (Issa et al., 2019):

$$\mathcal{L}(X \to Z \mid Y) \triangleq \sup_{S:\ S-(X,Y)-Z} I_\infty(S; Z \mid Y). \tag{9}$$

In comparison to the unconditional variant, this formalisation does not require $S \neq Y$. Therefore, it does not restrict the adversary's absolute gain from observing Z, but only restricts the leakage about sensitive attribute S *to its fundamental limit*, i.e., the leakage caused by the learning task label itself.

Interpretation. A feature map that satisfies the LPP in Definition 2 with a value of $\gamma \approx 0$ restricts the information available to the service provider to what is necessary for the intended purpose of the system: produce an accurate prediction of the task label Y. Hence, this formalisation supports the data protection principle of *purpose limitation* which states that "data should only be collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes" (European Parliament and Council of the European Union, 2016).

We must stress that, despite minimizing the information available to the service provider, this definition does not imply that a feature map that satisfies the LPP will necessarily prevent all harmful inferences. The LPP only limits the risk of revealing $Z = f_E(X)$ to the fundamental risk already incurred by revealing Y itself. In Section 4.1, we empirically show that even the fundamental leakage can lead to inferences that might violate users' expectations about the information they reveal through the use of a service—a violation of contextual integrity (Nissenbaum, 2004)—and lead to harms (Citron & Solove, 2022).

Perfect LPP. We first analyse the feasibility of *perfect LPP* with $\gamma = 0$. Perfect LPP implies that the representation Z leaks strictly no more information about the data X than already revealed by the task label Y itself. As we show next, perfect LPP is possible only under a restrictive assumption on the data distribution.

Proposition 1. *There exists a feature map $f_E(X)$ that fulfils the LPP with $\gamma = 0$ if and only if we have a Markov chain $X - Y - Z$.*

One class of data distributions that fulfils this condition and where perfect LPP is feasible is any distribution in which the intended task label Y is a deterministic function of the input data X (see, e.g., Fischer (2020)):

Corollary 1. *If $Y = g(X)$ for some function $g : \mathbb{X} \to \mathbb{Y}$, there exists $f_E(X)$ which achieves perfect LPP with $\gamma = 0$.*

As we discuss in Section 2, in many realistic scenarios, however, there is inherent uncertainty about an example's true label due to, for instance, label noise. Under Assumption A, perfect LPP is unattainable by non-trivial feature maps:

Corollary 2. *Under Assumption A, the feature map $f_E(X)$ satisfies LPP with $\gamma = 0$ if and only if it is fully random or constant: $f_E(X) \perp\!\!\!\perp X$.*

LPP and Utility Trade-Off. We now study whether we can find a feature map that achieves the LPP in Definition 2 with $\gamma > 0$, and simultaneously has good utility for the intended learning task. We show the general trade-off between a representation's utility and the LPP requirement:

Theorem 2 (LPP and Utility Trade-Off)**.** Suppose that $P_{Y|X}$ is strictly positive (Assumption A). Then, for $\alpha \in \{1, \infty\}$, the following two properties cannot hold at the same time:

I) $Z = f_E(X)$ satisfies the LPP with parameter γ

II) $I_\alpha(Y; Z) > \gamma$

See Appendix A for a full proof and Fig. 2 for an illustration.

One way to see why the trade-off holds is through the lens of attributes that reveal maximum possible information about X from Z (see Appendix A for a formal description). Intuitively, we would expect that by allowing for fundamental leakage, i.e., by conditioning on Y, we might reduce maximal leakage in those cases where Y is a maximally revealing attribute. However, as we show in Appendix A, strict positivity of the posterior distribution (Assumption A) prevents the task label Y from being a maximally revealing attribute, thus the least-privilege and utility trade-off remains the same as under unconditional LPP (Theorem 1).

Related Formalisms. Our formalisation of the LPP is closely related to the Minimum Necessary Information (MNI) principle introduced by Fischer (2020). The MNI criterion demands that a representation Z should maximise the mutual information $I(Y; Z)$ while minimising the conditional information $I(X; Z \mid Y)$. The LPP and MNI differ in their leakage measure, but are equal in the optimal point of $\gamma = 0$ in which $I(X; Z \mid Y) = \mathcal{L}(X \to Z \mid Y) = 0$.

The question of *finding* a feature representation that maximises utility but satisfies the LPP posed by Theorem 2 can be seen as a variant of the conditional entropy bottleneck (CEB) problem (Fischer, 2020), which, in turn is a variant of the standard information bottleneck problem (Asoodeh & Calmon, 2020; Tishby et al., 2000). The CEB problem aims to recover a representation that satisfies the MNI principle.

Unlike in the MNI and CEB, we require a leakage measure that captures the original claim from prior work that

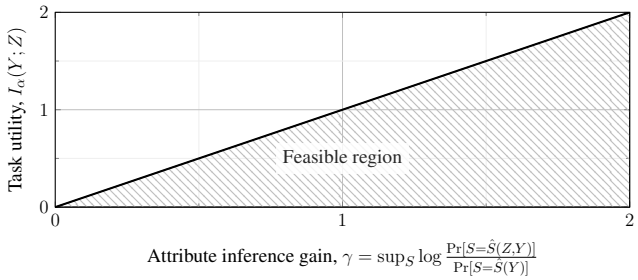

Figure 2: γ-LPP limits maximum utility $I_\alpha(Y;Z)$ of a representation Z to the greyed-out region.

it is possible to learn representations that protect against *any* harmful inferences, i.e., protect even against *worst-case inference* adversaries. Maximal leakage satisfies this requirement and in addition has a direct operational meaning in terms of inference gain (see Eq. (4)).

Other works, such as Mireshghallah et al. (2021) or Maeng et al. (2024), differ not only in their leakage measure but crucially do not consider a task's fundamental leakage, i.e., do not define the trade-off in terms of the *conditional* leakage $I_\alpha(S, Z \mid Y)$. In addition, Mireshghallah et al. (2021) and other formalisations of related information-theoretic problems, such as the Privacy Funnel (Makhdoumi et al., 2014; Asoodeh & Calmon, 2020), only consider a fixed sensitive attribute which does not adequately capture the least-privilege requirement.

LPP vs. LDP. Melis et al. (2019) suggest that LPL might be used to limit unintended information leakage as an alternative approach to record-level differential privacy (Dwork et al., 2014) which they find imposes a high utility cost that prevents its application in many realistic MLaaS scenarios. The hope is that LPL might provide a better trade-off between utility and restricting unintended information leakage than differental privacy. Our formalisation of the LPP enables us to formally analyse this claim.

First, let us define local differential privacy (LDP), a variant of differential privacy that is relevant to our MLaaS setting (see Fig. 1).

Definition 3 (LDP)**.** *A feature map $f_E(X) = Z$ satisfies ε-LDP if for any $x, x' \in \mathbb{X}$, and any $z \in \mathbb{Z}$, we have:*

$$\frac{\Pr[f_E(x) = z]}{\Pr[f_E(x') = z]} \leq 2^\varepsilon. \tag{10}$$

Conventionally, a power of e instead of 2 is used, but we use the latter for consistency with our previous definitions.

Given this definition, we show that LDP implies LPP:

Proposition 2. *A feature map that satisfies ε-LDP also satisfies ε-LPP.*

An important implication is that *a feature map that satisfies the LPP comes with the same utility trade-off as one that satisfies LDP*. Thus, in the setting of Theorem 2, LPL cannot provide an alternative to differential privacy at a lower utility cost as envisioned by prior work (Osia et al., 2018; Melis et al., 2019).

Takeaways. In summary, Theorem 2 implies that in many realistic applications (see Assumption A) there is a stringent trade-off between a representation's utility and achieving the LPP. Notably, this holds for *any* feature representation *regardless of the learning technique used to obtain the feature mapping or the exact model architecture*. In the next section, we show the practical implications of this result.

4. Empirical Evaluation

In this section, we empirically validate our theoretical results. We demonstrate that the fundamental trade-off between a representation's utility for its intended task and the LPP applies to *any feature representation regardless of the feature learning technique, model architecture, or dataset*. Due to space constraints, in the main body of the paper, we only present results for an image dataset, two different learning techniques, and a single model architecture. We discuss additional results that confirm that our theoretical results hold on a wider range of learning techniques, models, and datasets in Section 4.2, with their details deferred to Appendix B.

Data. In our main experiment, we use the LFWA+ image dataset which has multiple binary attribute labels for each image (Huang et al., 2008). The full dataset contains $13,143$ examples which we split in the following way: 20% of records are given to the adversary as an auxiliary dataset D_A. The remaining $10,514$ records are split $80/20\%$ across a train D_T and evaluation set D_E.

Model. We choose a deep convolutional neural network (CNN256) used by Melis et al. (2019), the first work to propose feature learning under a least-privilege principle. In the main results, we use the network's last layer representation as feature map $Z = f_E(X)$. We choose the last layer because, as Melis et al. (2019) and Mo et al. (2021) show, higher layer representations are expected to be more learning task specific and hence should restrict leakage of data attributes other than the task more than lower layer representations. In Appendix B, we include experiments with representations from other hidden layers of the model and for a MLaaS training setting in which users share gradients from all layers of the model.

Adversaries. To evaluate leakage of a given attribute, we instantiate Bayes-optimal adversaries with access to the auxiliary set D_A of labelled examples $\mathbf{r}_i = (x_i, y_i, s_i)$ for

$i = 1, \ldots, A$. The label-only adversary $\hat{S}(Y)$ computes the relative frequency counts over D_A to estimate:

$$\tilde{\Pr}[S = s \mid Y = y] \triangleq \frac{\sum_{i=1}^{A} \mathbf{1}[s_i = s, y_i = y]}{\sum_{i=1}^{A} \mathbf{1}[y_i = y]},$$

and outputs a guess according to:

$$\hat{S}(Y = y) = \arg\max_{s} \tilde{\Pr}[S = s \mid Y = y].$$

The features adversary $\hat{S}(Z, Y)$ is given black-box access to the feature mapping of the trained model. To collect a train set, the features adversary submits each example $x_i \in D_A$ to the model and receives back its representation $z_i = f_E(x_i)$. The adversary then trains a Random Forest classifier with 50 decision trees on the collected samples (z_i, y_i, s_i) to estimate $\tilde{\Pr}[S = s \mid Z = z, Y = y]$. We opt for Random Forest as an attack model based on its superior performance over other classifiers we tested.

Experimental Setup. In each experiment, we select one out of 12 attributes from the LFWA+ dataset as the model's learning task Y and a second attribute as the sensitive attribute S targeted by the adversary. We select attributes for which we expect the distribution $P_{Y|X}$ to be strictly positive due to their subjective nature (see Assumption A). We repeat each experiment 5 times to capture randomness of our measurements for both the model and adversary, and show average results across all 5 repetitions. At the start of the experiment, we split the data into train D_T, evaluation D_E, and auxiliary set D_A. We train the model CNN256 on the train set D_T for the chosen learning task and then estimate its utility on the evaluation set D_E. We estimate the model's utility as its multiplicative gain in task accuracy $\tilde{I}_\infty(Y; Z)$ (see Eq. (1)). After model training and evaluation, we train both the label-only and features adversary on the auxiliary data D_A. For a given sensitive attribute S, we estimate the adversary's gain as $\tilde{I}_\infty(S; Z \mid Y)$ (see Eq. (4)).

Learning Techniques. We implement two learning techniques: (1) standard ERM with SGD, and (2) attribute censoring to learn representations that hide a given sensitive attribute. In our main experiment, we use the gradient reversal strategy (GRAD) introduced by Raff & Sylvester (2018) for censoring with a censoring parameter of 100. We choose GRAD because it effectively hides the chosen sensitive attribute without large drops in model performance (Zhao et al., 2020), and, unlike other censoring techniques, can be applied to any model architecture. In Appendix B, we show equivalent results for another learning technique that aims to hide sensitive information, adversarial representation learning.

4.1. The Potential Harms of Fundamental Leakage

Fig. 3 (*left*) shows the adversary's gain $\tilde{I}_\infty(S; Y)$ in predicting sensitive attribute S from Y for pair-wise combinations of 12 learning tasks and sensitive attributes. This gain represents the fundamental leakage of a system that fulfils its intended purpose, i.e., that produces accurate task labels for the chosen learning task. The matrix in Fig. 3 (*right*) shows the absolute pairwise Pearson's correlation coefficient between attribute labels. Fig. 3 shows that, as expected, a strong linear relationship between the learning task and the sensitive attribute targeted by the adversary leads to a large fundamental leakage. For instance, 'Attractive' and 'Male' are negatively correlated with a correlation coefficient of -0.3094. This is already enough to increase the adversary's gain in inferring sensitive attribute 'Male' when the learning task is 'Attractive'. Although the increase is small in this case, it illustrates how inferences due to a task's fundamental leakage can not only be counterintuitive, but also reveal information that could lead to harm (in this case, discrimination due to gender).

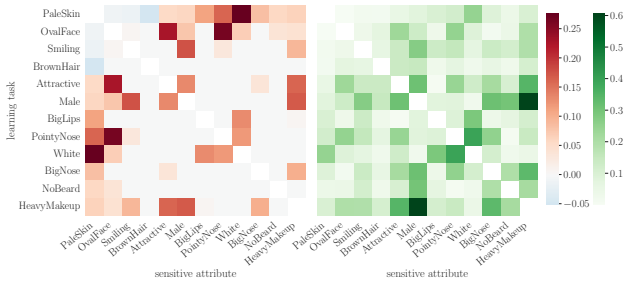

Figure 3: **Fundamental leakage: the task label reveals information about other data attributes, which might not be obvious to data subjects.** Attribute inference gain of the label-only adversary (*left*) and pairwise Pearson's correlation between attributes (*right*). In the LFWA+ dataset, the 'Attractive' label is highly correlated with the perceived gender. Thus, predicting the 'Attractive' label will reveal information about gender.

The fundamental leakage for some pairs of tasks and sensitive attributes is highly concerning. It implies that users reveal to the service provider not only their task label but also *any attribute that is correlated with the chosen task*. This consequence is rarely made explicit to users when they are informed about the data collection and processing, and might lead to unexpected harms beyond those associated with revealing the task label itself. In the example above, a user expecting to only reveal 'attractiveness' might not expect that their gender is revealed, with the ensuing risks of discrimination. To better inform users about their risk, providers would have to list all sensitive attributes that might be leaked through a task's fundamental leakage. Knowing *all* such attributes is infeasible. To address this problem, service providers could empirically evaluate whether attributes considered particularly sensitive are part of the fundamental leakage and inform data subjects about the result.

4.2. The Least-Privilege And Utility Trade-Off

One way to interpret Theorem 2 is that whenever the features learned by a model are useful for a given prediction task, there always exists a sensitive attribute for which an adversary gains an advantage from observing a target record's feature representation. We experimentally show this fundamental limit of least-privilege learning.

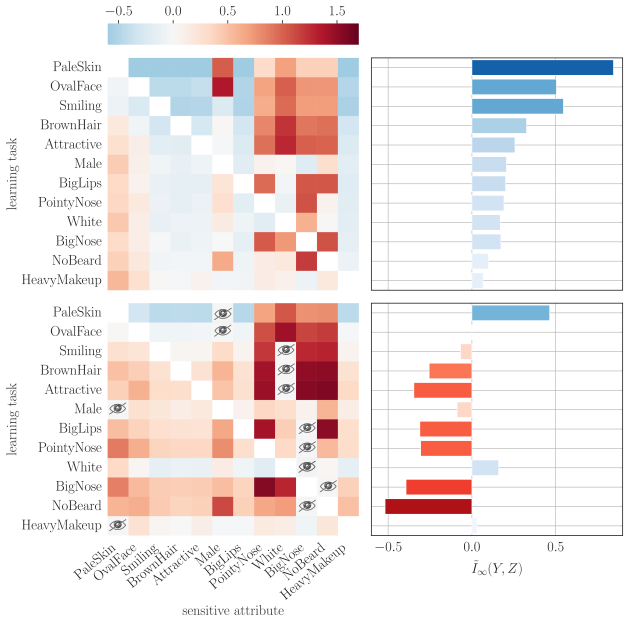

Figure 4: **If the model-generated representations have utility for the task (*right*), there exists a sensitive attribute with an even higher inference gain for the adversary (*left, red means more leakage*).** This holds for both standard ERM (*top*) and attribute censoring (*bottom*) where we censor the attribute with highest leakage in the respective ERM model (marked as 👁). Censoring has a 'whack-a-mole' effect: as we censor one attribute, leakage of another attribute increases.

Fig. 4 compares the trade-off between utility and attribute leakage of models trained with standard SGD (*top*) and with GRAD to censor the representation of a single sensitive attribute (*bottom*). The blue bars in Fig. 4 (*right*) show the model's utility for learning task Y measured as $\tilde{I}_\infty(Y;Z)$. The heatmaps in Fig. 4 (*left*) show the difference between the adversary's inference gain and the model's utility:

$$\Delta_{\mathsf{Adv}} \triangleq \tilde{I}_\infty(S;Z \mid Y) - \tilde{I}_\infty(Y;Z).$$

Each row corresponds to a learning task Y, and each column to a sensitive attribute targeted by the adversary.

Across Learning Tasks. In Fig. 4 (*top*), we see that in the LFWA+ dataset the features learned by a model that predicts attribute 'Smiling' increase the adversary's inference gain for attribute 'White'. Different tasks result in a high leakage for different attributes, e.g., 'OvalFace' reveals a lot of additional information about gender, while 'HeavyMakeup' is indicative of 'PaleSkin' and 'Attractive' but does not have much influence on the inference power of other attributes. Importantly, for *every* task, there is at least one sensitive attribute with $\Delta_{\mathsf{Adv}} > 0$. These results confirm that the features learned by a model trained to perform well on its learning task reveal more information than a task's fundamental leakage and thus violate the LPP. Fig. 5 in Appendix B shows these results in terms of the adversary's absolute inference gain $\tilde{I}_\infty(S;Z \mid Y)$. It confirms that the adversary's absolute inference gain increases with a model's performance on its intended task.

Across Learning Techniques. In Fig. 4 (*bottom*), we show results for models trained under attribute censoring, a common technique used to address unintended information leakage (Song & Shmatikov, 2019; Brown et al., 2022; Zhao et al., 2020). For each task, we censor the attribute with the highest leakage under standard training. First, as expected, censoring limits the leakage of the censored attribute. However, we observe that the trade-off from Theorem 2 holds: In all our experiments, an adversary can find another data attribute with $\Delta_{\mathsf{Adv}} > 0$ that thus violates the LPP.

In Fig. 6 in Appendix B, we present experiments for another learning technique, adversarial representation learning, that has been shown to be an efficient censoring technique (Zhao et al., 2020). We find similar results: For two out of the four sensitive attributes we test, the adversary's inference gain exceeds the features' utility $\tilde{I}_\infty(S,Z \mid Y) > \tilde{I}_\infty(Y,Z)$. These results confirm that the trade-off between a representation's utility and the LPP holds regardless of the learning technique used to obtain $f_E(X)$.

Across Hidden Layers. We conduct additional experiments that show that the strict trade-off of Theorem 2 applies regardless of the model architecture (see Fig. 9 in Appendix B) or the hidden layer at which the adversary observes a record's representation (see Fig. 8 in Appendix B). These results are in line with prior work by Mo et al. (2021) and Melis et al. (2019). Even though higher layers are expected to be more learning task specific, an adversary can misuse the shared representations at any layer to predict attributes other than the intended task. We show that this leakage even exceeds the fundamental leakage of the task label itself and violates the LPP.

Across Datasets and Model Architectures. Besides images, tabular data is another type of data for which representation sharing is commonly suggested as a technique to limit unintended information leakage (Zhao et al., 2020; Arik & Pfister, 2021; Creager et al., 2019; Raff & Sylvester, 2018). In Figs. 6 and 10 in Appendix B, we show that the trade-off

also holds for this type of data (and the associated model architectures) using two tabular datasets: it is not possible to find feature representations that have high utility for a prediction task but fulfil the LPP.

In Collaborative Learning Settings. The gradient updates shared with a ML service provider in the collaborative learning setting are a noisy version of the feature representations learned by the model (Melis et al., 2019). In Fig. 7 in Appendix B, we show that hence the fundamental trade-off of Theorem 2 equally applies to gradient updates shared for model training. Our experiments thus show that LPL cannot, as envisioned by Melis et al. (2019), provide the promised solution to address unintended feature leakage in the collaborative learning setting.

Takeaways. Our experiments confirm *the fundamental trade-off* we derive in Theorem 2: the representations generated by models that perform well on their intended task fail to fulfil the LPP. Censoring techniques can be used to limit the adversary's inference gain on a particular attribute but cannot avoid the trade-off: there always exists an attribute, other than the learning task, that violates the LPP. This 'whack-a-mole' effect is a phenomenon observed in related scenarios, such as, privacy-preserving data publishing (Narayanan & Shmatikov, 2019). Our experiments are the first to study the trade-off between a representation's utility and unintended information leakage across a large combination of learning tasks and sensitive attributes. In contrast to prior work that evaluates unintended feature leakage for a single learning task and fixed sensitive attribute, this enables us to show that although it is sometimes possible to restrict information leakage for a single attribute, the learned representations do not fulfil the LPP.

5. Limitations and Future Work

Our results hold under a mild assumption on the data distribution (Assumption A) and for a specific operational measure of leakage (Section 3.2). By relaxing these two aspects, it might be possible to achieve a more favourable trade-off than what we find.

First, the trade-off could possibly be less stringent for specific data distributions for which Assumption A does not hold. In such cases, it might be possible that for some learning tasks, there exists a representation that provides some utility gain and at the same time constrains the maximal leakage, i.e., that simultaneously fulfils the two conditions stated in Theorem 2. It could be a direction for future work to formally analyse whether such representations are achievable for distributions that do not satisfy Assumption A.

Second, the leakage measure used to formalise the LPP could be relaxed by restricting the set of possible inference targets to a less rich class of distributions. For instance, the class of attributes targeted by the adversary could be limited to those "efficiently identifiable" (Hébert-Johnson et al., 2018). However, such modifications would require a different analytical toolbox and deviate from the promise of LPL to leak "nothing else" (Brown et al., 2022) but the intended task. In addition, although the leakage measure considered here might seem strong, it is already significantly weaker than standard privacy notions, such as LDP, as it measures attribute inference success in expectation over the population as opposed to worst-case risks.

6. Conclusions

The promise of least-privilege learning — to learn feature representations that are useful for a given task but avoid leakage of any information that might be misused and cause harm — is extremely appealing. In this paper, we show that in certain realistic settings where there exists inherent uncertainty about the task labels, any representation that provides utility for its intended task must always leak information about attributes other than the task and thus does neither fulfil the least-privilege principle nor provide a better utility-privacy trade-off than existing formal privacy-preserving techniques. Furthermore, we show that, due to a task's fundamental leakage, even representations that fulfil the least-privilege principle might not prevent all the potential harms. These issues apply to any setting in which users share representations instead of raw data records and regardless of the learning technique used to learn the feature mapping that produce these representations.

Impact Statement

This paper validates claims from prior works that the concept of least-privilege learning might provide a promising avenue to mitigate harms of potential misuse of machine learning. Prior works repeatedly suggest that least-privilege learning might achieve a better trade-off between revealing information about an intended task and unintended information leakage than existing techniques, such as differential privacy. In this paper, we formally prove that least-privilege learning might not fulfil this promise: We show that there is an inherent trade-off between a representation's utility and its unintended information leakage that cannot be overcome by any learning technique. The results of this paper will help practitioners to evaluate the potential benefits and shortcomings of a newly proposed technique. Our formalisation of a previously only informally stated property will enable researchers to formally compare the trade-offs of alternative methods to prevent data misuse in machine learning.

Acknowledgements

This work is partially funded by the Swiss National Science Foundation under grant 200021-188824. NP acknowledges support from the Canada CIFAR AI Chair. We would like to thank Christian Knabenhans for providing a basis for the implementation of the empirical evaluation of this work.

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2016.

Arik, S. Ö. and Pfister, T. Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Arimoto, S. Information measures and capacity of order α for discrete memoryless channels. *Topics in information theory*, 1977.

Asoodeh, S. and Calmon, F. P. Bottleneck problems: An information and estimation-theoretic view. *Entropy*, 22 (11):1325, 2020.

Boenisch, F., Dziedzic, A., Schuster, R., Shamsabadi, A. S., Shumailov, I., and Papernot, N. Reconstructing individual data points in federated learning hardened with differential privacy and secure aggregation. In *IEEE S&P*, 2023.

Brown, G., Martinez-del Rincon, J., and Miller, P. Least privilege learning for attribute obfuscation. In *Pattern Recognition*, 2022.

Chatzikokolakis, K., Cherubin, G., Palamidessi, C., and Troncoso, C. Bayes security: A not so average metric. In *36th IEEE Computer Security Foundations Symposium, CSF 2023, Dubrovnik, Croatia, July 10-14, 2023*, pp. 388–406. IEEE, 2023.

Chi, J., Owusu, E., Yin, X., Yu, T., Chan, W., Tague, P., and Tian, Y. Privacy partitioning: Protecting user data during the deep learning inference phase. *arXiv preprint arXiv:1812.02863*, 2018.

Citron, D. K. and Solove, D. J. Privacy harms. *BUL Rev.*, 102:793, 2022.

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. Flexibly fair representation learning by disentanglement. In *ICML*, 2019.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN, 2016.

Fischer, I. The conditional entropy bottleneck. *Entropy*, 22 (9):999, 2020.

Ganju, K., Wang, Q., Yang, W., Gunter, C. A., and Borisov, N. Property inference attacks on fully connected neural networks using permutation invariant representations. In *ACM SIGSAC Conference on Computer and Communications Security*, 2018.

Hao, K. How Apple personalizes Siri without hoovering up your data. https://www.technologyreview.com/2019/12/11/131629/apple-ai-personalizes-siri-federated-learning/, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *arXiv*, 2015.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *ICML*, 2018.

Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Issa, I., Wagner, A. B., and Kamath, S. An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657, 2019.

Kohavi, R. and Becker, B. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Adult, 2013.

Liao, J., Kosut, O., Sankar, L., and du Pin Calmon, F. Tunable measures for information leakage and applications to privacy-utility tradeoffs. *IEEE Transactions on Information Theory*, 2019.

Maeng, K., Guo, C., Kariyappa, S., and Suh, G. E. Bounding the invertibility of privacy-preserving instance encoding using fisher information. In *NeurIPS*, 2024.

Makhdoumi, A., Salamatian, S., Fawaz, N., and Médard, M. From the information bottleneck to the privacy funnel. In *2014 IEEE Information Theory Workshop (ITW 2014)*, pp. 501–505. IEEE, 2014.

McMahan, B. and Ramage, D. Federated learning: Collaborative machine learning without centralized training data. http://ai.googleblog.com/2017/04/federated-learning-collaborative.html, 2017.

Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting Unintended Feature Leakage in Collaborative Learning. In *SP*, 2019.

Mireshghallah, F., Taram, M., Jalali, A., Elthakeb, A. T. T., Tullsen, D., and Esmaeilzadeh, H. Not all features are equal: Discovering essential features for preserving prediction privacy. In *Proceedings of the Web Conference 2021*, 2021.

Mo, F., Borovykh, A., Malekzadeh, M., Haddadi, H., and Demetriou, S. Layer-wise Characterization of Latent Information Leakage in Federated Learning. *arXiv:2010.08762*, 2021.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets: a decade later. https://www.cs.princeton.edu/~arvindn/publications/de-anonymization-retrospective.pdf, 2019.

Nissenbaum, H. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

Osia, S. A., Taheri, A., Shamsabadi, A. S., Katevas, K., Haddadi, H., and Rabiee, H. R. Deep private-feature extraction. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

Raff, E. and Sylvester, J. Gradient Reversal Against Discrimination. In *FACCT*, 2018.

Raji, I. D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. Ai and the everything in the whole wide world benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., and Jégou, H. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research. PMLR, 2019.

Saltzer, J. H. and Schroeder, M. D. The protection of information in computer systems. *Proceedings of the IEEE*, 1975.

Song, C. and Shmatikov, V. Overlearning Reveals Sensitive Attributes. In *NIPS*, 2019.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Texas Department of State Health Services, Austin, Texas. Texas Hospital Inpatient Discharge Public Use Data File 2013 Q1-Q4. https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm, 2013. Accessed 2020-06-01.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Wang, J., Zhang, J., Bao, W., Zhu, X., Cao, B., and Yu, P. S. Not just privacy: Improving performance of private deep learning in mobile cloud. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

Zhao, H., Chi, J., Tian, Y., and Gordon, G. J. Trade-offs and guarantees of adversarial representation learning for information obfuscation. *NeurIPS*, 2020.

A. Formal Details

Maximal Leakage. First, let us provide some useful properties of maximal leakage.

Maximal leakage and conditional maximal leakage are defined (Issa et al., 2019) as follows:

$$\mathcal{L}(X \to Z) \triangleq \sup_{S:\ S-X-Z} I_\infty(S;Z), \qquad \mathcal{L}(X \to Z \mid Y) \triangleq \sup_{S:\ S-(X,Y)-Z} I_\infty(S;Z \mid Y). \tag{11}$$

Lemma 1 (Issa et al. (2019))**.** *Maximal leakage has the following properties:*

I) Maximal leakage bounds mutual information for $\alpha \in \{1, \infty\}$: $\mathcal{L}(X \to Z) \geq I_\alpha(X;Z)$

II) Maximal leakage satisfies data processing inequalities in a Markov chain $Y - X - Z$:

$$\mathcal{L}(Y \to Z) \leq \min\{\mathcal{L}(X \to Z), \mathcal{L}(Y \to X)\}$$

III) Maximal leakage and conditional maximal leakage have the following closed forms:

$$\mathcal{L}(X \to Z) = \log\left(\sum_{z\in\mathbb{Z}} \max_{x\in\mathbb{X}} P(z \mid x)\right)$$

$$\mathcal{L}(X \to Z \mid Y) = \log\left(\max_{y\in\mathbb{Y}} \sum_{z\in\mathbb{Z}} \max_{x\in\mathrm{supp}(X|Y=y)} P(z \mid x, y)\right)$$

Maximally Revealing Attributes. We demonstrate that there exists an attribute that is maximally revealing, i.e., achieves the supremum $\sup_{S:\ S-X-Z} I_\infty(S;Z)$, and show how to construct this attribute.

Definition 4. *For a given P_X, a* maximally revealing attribute *S^* (see "shattering distribution", Issa et al., 2019) is defined over $\mathbb{S} \triangleq \mathbb{X} \times \mathbb{Z}$ via the following conditional probability distribution:*

$$P_{S^*|X}((x',k) \mid x) \triangleq \begin{cases} \frac{p_{\min}}{P_X(x)}, & x' = x \text{ and } 1 \leq k \leq \lfloor r(x) \rfloor \\ 1 - \frac{(\lceil r(x) \rceil - 1)\cdot p_{\min}}{P_X(x)}, & x' = x \text{ and } k = \lceil r(x) \rceil \\ 0, & x' \neq x, \end{cases} \tag{12}$$

for any $x, x' \in \mathbb{X}$, and $k \in [1, \ldots, \lceil r(x) \rceil]$, where $p_{\min} \triangleq \min_{x\in\mathbb{X}} P_X(x)$ and $r(x) \triangleq {}^{P_X(x)}/_{p_{\min}}$.

Proposition 3 (Issa et al. (2019))**.** *For any $Z = f_E(X)$ and P_X, the maximally revealing attribute S^* leads to the highest inference gain:*

$$\sup_{S:\ S-X-Z} I_\infty(S;Z) = I_\infty(S^*;Z) \tag{13}$$

Proof. It is sufficient to show that for any $P_{Z|X}$ and P_X, we have $\mathcal{L}(X \to Z) = I_\infty(S^*;Z)$. By Lemma 1 and the definition of I_∞, this is equivalent to showing:

$$\log\left(\sum_{z\in\mathbb{Z}} \max_{x\in\mathbb{X}} P_{Z|X}(z \mid x)\right) = \log\left(\frac{\sum_{z\in\mathbb{Z}} \max_{s\in\mathbb{S}} P_{S^*,Z}(s,z)}{\max_{s\in\mathbb{S}} P_{S^*}(s)}\right). \tag{14}$$

We can see that $\max_{s\in\mathbb{S}} P_{S^*|Z}(s,z)$ from the RHS has the following form:

$$\max_{(x',k)\in\mathbb{S}} P_{S^*,Z}((x',k),z) = \max_{(x',k)\in\mathbb{S}} P_{S^*|X}((x',k) \mid x') \cdot P_X(x') \cdot P_{Z|X}(z \mid x') \tag{15}$$

$$= \max_{(x',k)\in\mathbb{S} \text{ s.t. } k=1} P_{S^*|X}((x',1) \mid x') \cdot P_X(x') \cdot P_{Z|X}(z \mid x') \tag{16}$$

$$= \max_{x\in\mathbb{X}} \frac{p_{\min}}{P_X(x)} \cdot P_X(x) \cdot P_{Z|X}(z \mid x) = p_{\min} \cdot \max_{x\in\mathbb{X}} P_{Z|X}(z \mid x), \tag{17}$$

where the first equality is by Markov chain $S^* - X - Z$, and the second and third are by Definition 4. Combining the resulting expression and the fact that by Definition 4 we have

$$\max_{s\in\mathbb{S}} P_{S^*}(s) = \max_{s\in\mathbb{S}} \sum_{x\in\mathbb{X}} P_{S^*|X}(s)\cdot P_X(x) = p_{\min}, \tag{18}$$

we get the sought result. □

We also show the following technical property of a maximally revealing attribute.

Proposition 4. *Suppose that $|\mathbb{X}| > 1$. The conditional distribution $P_{S^*|X}$ is non-positive: there exist a $s \in \mathbb{S}$ and $x \in \mathbb{X}$ such that $P_{S^*|X}(s \mid x) = 0$.*

Proof. By Definition 4, for any $x \in \mathbb{X}$ it is sufficient to take any $(x', k) \in \mathbb{S}$ where $x' \neq x$.

A.1. Omitted Proofs

Next, we provide the proofs of the formal statements in the main body of the paper.

Proof of Theorem 1. By construction, we have a Markov chain $Y - X - Z$. Observe that $S^* \neq Y$ by Proposition 4 and strict positivity (Assumption A). Therefore,

$$\sup_{S:\ S-X-Z,\ S\neq Y} I_\infty(S;Z) = I_\infty(S^*;Z) = \sup_{S:\ S-X-Z} I_\infty(S;Z) = \mathcal{L}(X\to Z) \tag{19}$$

by Proposition 3 and the definition of maximal leakage. Finally, applying the properties in Lemma 1, we have $I_\alpha(Y;Z) \leq \mathcal{L}(Y\to Z) \leq \mathcal{L}(X\to Z) = \sup_{S:\ S-X-Z,\ S\neq Y} I_\infty(S;Z) \leq \gamma$. □

Remark 1. *Theorem 1 only requires that $Y \neq S^*$. This holds under weaker assumptions than the strict positivity of the posterior (Assumption A). For instance, it is sufficient that there exist any x, x', y such that both $P_{X|Y}(x \mid y) > 0$ and $P_{X|Y}(x' \mid y) > 0$.*

Proof of Proposition 1. For $\gamma = 0$, by Lemma 1 we have that $\mathcal{L}(X\to Z \mid Y) = I(X;Z \mid Y) = 0$. As conditional mutual information is consistent with conditional independence, this holds if and only if we have Markov chain $X - Y - Z$. □

Proof of Corollary 1. If $Y = g(X)$, where $g(\cdot)$ is a deterministic function, we can choose $Z = f(Y) = f(g(X))$. Evidently, this satisfies the Markov relationship $f(g(X)) - g(X) - X$, *i.e.,* $Z - Y - X$. (Note that $f(\cdot)$ may even be a randomized function for this to hold.) But this Markov relation implies $\mathcal{L}(X\to Z \mid Y) = 0$. □

Corollary 2 follows as a consequence of the conditional independence in Proposition 1 and the fact that we have $Y - X - Z$ by construction.

Proof of Theorem 2. First, we show that in our Markov chain and under the strictly positive posterior assumption, we have a surprising result that $\mathcal{L}(X\to Z \mid Y) = \mathcal{L}(X\to Z)$.

To see this, observe that maximal leakage has the following closed form by Lemma 1:

$$\mathcal{L}(X\to Z \mid Y) = \log\left(\max_{y\in\mathbb{Y})}\sum_{z\in\mathbb{Z}} \max_{x\in\mathrm{supp}(X|Y=y)} P(z \mid x, y)\right) \tag{20}$$

$$= \log\left(\max_{y\in\mathbb{Y}}\sum_{z\in\mathbb{Z}} \max_{x\in\mathrm{supp}(X|Y=y)} P(z \mid x)\right), \tag{21}$$

where the second equality is by the Markov chain $Y - X - Z$.

Next, observe that we have $P(x \mid y) \propto P(y \mid x) \cdot P(x)$ which, by assumption, is positive so long as $x \in \text{supp}(X)$. As a consequence, the support of X is independent of Y: $\text{supp}(X \mid Y = y) = \text{supp}(X)$, for any $y \in \mathbb{Y}$. Therefore, we can simplify the last form:

$$\mathcal{L}(X \to Z \mid Y) = \log \left(\max_{y \in \mathbb{Y}} \sum_{z \in \mathbb{Z}} \max_{x \in \text{supp}(X \mid Y = y)} P(z \mid x) \right) \tag{22}$$

$$= \log \left(\sum_{z \in \mathbb{Z}} \max_{x \in \mathbb{X}} P(z \mid x) \right), \tag{23}$$

which is an equivalent form of $\mathcal{L}(X \to Z)$.

Finally, to obtain the trade-off, it suffices to observe that $I_\alpha(Y; Z) \leq \mathcal{L}(Y \to Z) \leq \mathcal{L}(X \to Z) \leq \gamma$, where the the first and the second inequalities are by properties of maximal leakage in Lemma 1. □

One way to interpret this result is that the maximally revealing attribute S^* is not sensitive to conditioning on Y so long as we have strict positivity of the posterior (Assumption A).

Proof of Proposition 2. Issa et al. (2019) show that if the map $f_E(\cdot)$ is ε-LDP, then $\mathcal{L}(X \to Z) \leq \varepsilon$, where $Z = f_E(X)$. From the form in Eq. (22), we have that $\mathcal{L}(X \to Z \mid Y) \leq \mathcal{L}(X \to Z) \leq \varepsilon$. □

B. Empirical Evaluation of the Utility and LPP Trade-off

In this section, we provide additional details, as well as, additional empirical results that show that the strict trade-off between utility and the LPP hold for any feature representation regardless of the feature learning technique, model architecture, or dataset.

B.1. Details of Setup in Section 4

Following Melis et al. (2019), we use a convolutional network with three spatial convolution layers with 32, 64, and 128 filters, kernel size set to (3, 3), max pooling layers with pooling size set to 2, followed by two fully connected layers of size 256 and 2. We use ReLU as the activation function for all layers.

B.2. Adversarial Inference Gain

Fig. 5 shows the adversary's inference gain $I_\infty(S; Z \mid Y)$ which measures the adversary's classification accuracy for sensitive attribute S (each column) normalised by a task's (each row) fundamental leakage.

B.3. Across Learning Techniques

Theorem 2 implies that the strict trade-off between a representation's utility for its intended task and the LPP holds regardless of the learning technique used to obtain the feature map $f_E(X) = Z$. In Section 4.2, we show that indeed even with attribute censoring through gradient reversal, an adversary can always find a data attribute for which $\Delta_{Adv} > 0$ and that thus violates the LPP. In this section, we experimentally demonstrate that the same applies to other learning techniques that aim to hide sensitive information about the original data X, such as adversarial representation learning.

Zhao et al. (2020) empirically compare the trade-off between hiding sensitive information and task accuracy of various attribute obfuscation algorithms. They find that together with gradient reversal, Maximum Entropy Adversarial Representation Learning (MAX-ENT) provides the best trade-off. We run a simple experiment on the Adult dataset (Kohavi & Becker, 2013), the data used by Zhao et al., that shows that the trade-off predicted by Theorem 2 also applies to the representations learned by a model trained under MAX-ENT.

Experiment Setup. We use the exact same model architecture and data as Zhao et al. (2020). We train the model to predict attribute 'income' and adversaries for four sensitive attributes ('age', 'education', 'race', and 'sex'). We then calculate the utility and inference gain as described in Section 4.

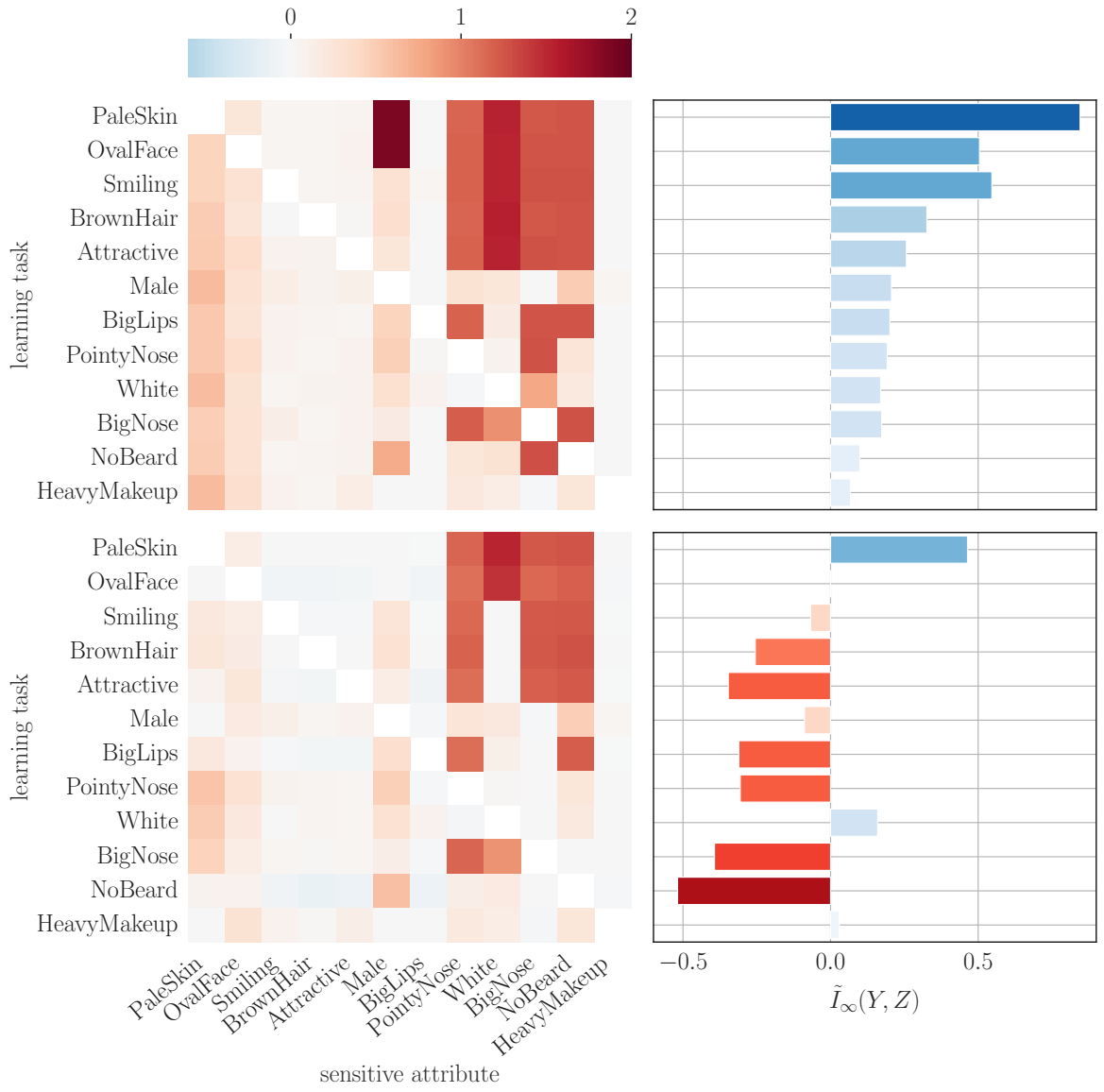

Figure 5: **If the model-generated representations have utility for the task (*right*), there exists a sensitive attribute with an even higher inference gain for the adversary (*left, red means more leakage*).** This holds for both standard ERM (*top*) and attribute censoring (*bottom*) where we censor the attribute with highest leakage in the respective ERM model (marked as 👁). Censoring has a 'whack-a-mole' effect: as we censor one attribute, leakage of another attribute increases.

Fig. 6 shows that, as expected, even for a model trained under attribute obfuscation with MAX-ENT, the adversary's inference gain exceeds the model's utility gain for two out of the four sensitive attributes tested. This further supports our theoretical finding that the trade-off between LPP and utility for a prediction task of a representation applies *regardless of how these representations are learned*.

B.4. Across MLaaS settings

We show that, as discussed in Section 2, the fundamental trade-off of Theorem 2 equally applies to the collaborative learning setting in which users share gradients instead of raw data records to reduce unintended information leakage (Melis et al., 2019).

Experiment Setup. We replicate the single batch property inference attack described by Melis et al. (2019). In this setting,

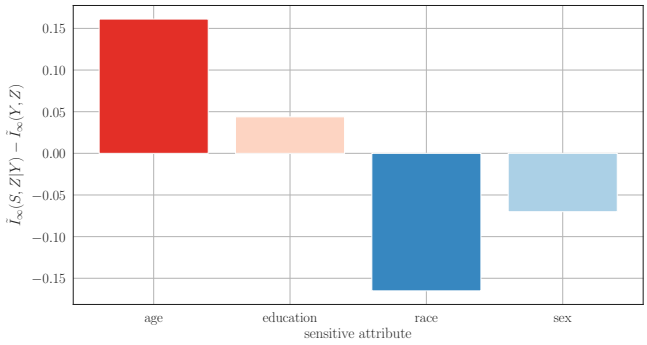

Figure 6: **The least-privilege and utility trade-off holds regardless of the representation learning technique used.** We show the delta between the adversary's inference gain and model utility for a model trained under MAX-ENT for learning task 'income' on the Adult dataset. For two out of four sensitive attributes tested, the adversary's inference gain exceeds the model's utility gain.

the feature representation $Z = f_E(X)$ shared with the service provider is a record's gradients computed across all layers of the model. As Melis et al. (2019) point out, these "gradient updates can [...] be used to infer feature values", i.e., are just a noisy version of the feature representations learned by the model.

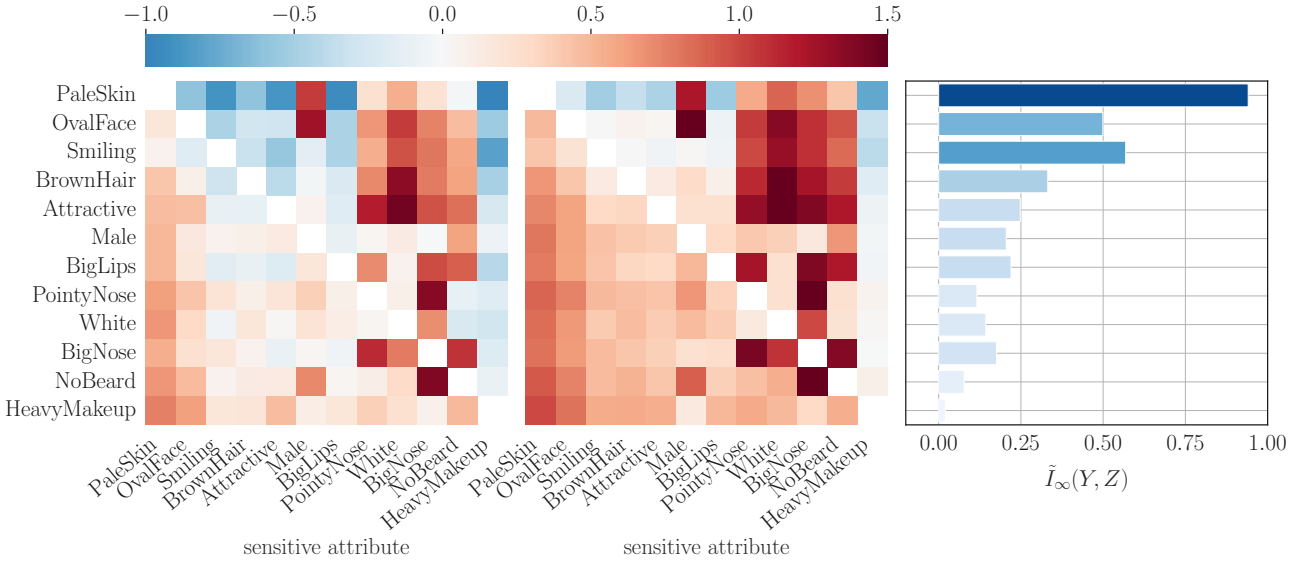

Figure 7: **The least-privilege and utility trade-off applies also the collaborative learning setting** The inference gain of an adversary that observes a record's representation in the form of gradients (*left*) or feature activations (*middle*) always violates the LPP for at least one sensitive attribute (red means that the adversary's inference gain is larger than the features' utility gain plotted on the (*right*))

Fig. 7 confirms that an adversary that only has indirect access to a record's feature representation in the form of gradients (*left*) can always find at least on sensitive attribute for which $\Delta_{\mathsf{Adv}} > 0$. The adversary's inference gain increases when the adversary has direct access to the feature representations (*middle*) but as predicted follows the fundamental trade-off predicted by Theorem 2 even when she only observes gradients.

B.5. Across Hidden Layers

To demonstrate that, as predicted by Theorem 2, the strict trade-off between features' utility for a downstream prediction task and the LPP applies regardless of a model's architecture, or the structure of the feature encoder $Z = f_E(X)$, we conduct additional experiments on the LFWA+ image dataset (see Section 4)

To show that the trade-off holds regardless of the structure of the feature encoder, we assess the leakage of the CNN256 model described in Section 4 across different hidden layers of the model. Fig. 8 shows the delta between the adversary's inference gain and the features' utility for the intended task $\Delta_{\mathsf{Adv}} = \tilde{I}_\infty(S, Z \mid Y) - \tilde{I}_\infty(Y, Z)$ across hidden layers of the model (from higher (*left*) to lower (*right*) layers). Following Theorem 2, for every learning task and representation there always exists at least one sensitive attribute for which $\Delta_{\mathsf{Adv}} > 0$ and that violates the LPP. In line with prior results (Melis et al., 2019; Mo et al., 2021), we find that lower layers (*right*) lead to a slightly higher inference gain for the adversary; although these differences are not significant.

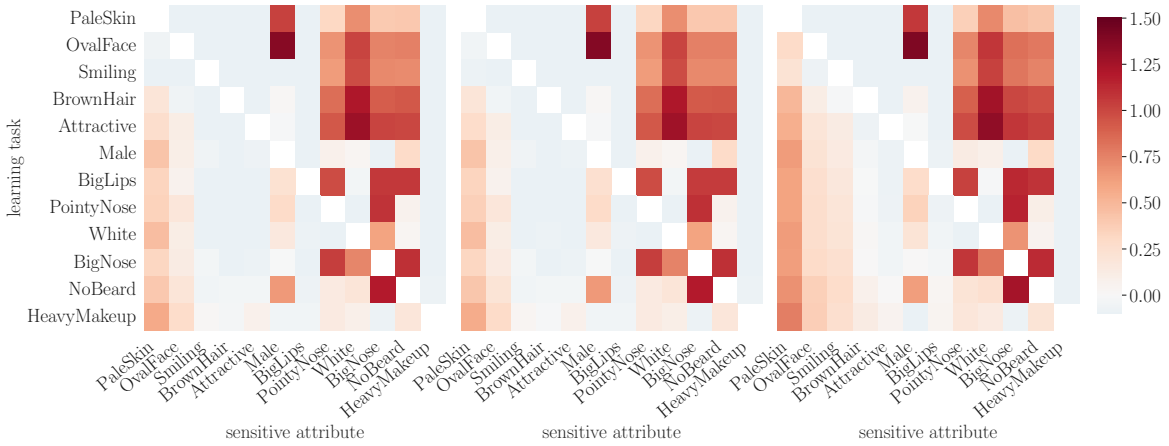

Figure 8: **The least-privilege and utility trade-off holds across hidden layers of the model** The inference gain of an adversary that observes a record's representation at the last (*left*), or first (*middle*) and second convolutional layer (*right*) of a CNN256 model, always violates the LPP for at least one sensitive attribute (red means that the adversary's inference gain is larger than the features' utility gain)

B.6. Across Model Architecture

To show that the trade-off holds regardless of the exact model architecture, we repeat the experiment outlined in Section 4 using the ResNet-18 model architecture from (He et al., 2015) implemented by PyTorch. Training batch size is 32, SGD learning rate is 0.01. The adversary is given access to the feature representations at the last hidden layer of the model.

Fig. 9 compares the trade-off between utility and attribute leakage of a CNN256 (*top*) and a ResNet18 (*bottom*) models, both trained with standard SGD. The blue horizontal bars in Fig. 9 (*right*) show the model's utility for learning task Y measured as $\tilde{I}_\infty(Y, Z)$. The heatmaps in Fig. 9 (*left*) show the difference between the adversary's inference gain and the model's utility $\Delta_{\mathsf{Adv}} = \tilde{I}_\infty(S, Z \mid Y) - \tilde{I}_\infty(Y, Z)$. Each row corresponds to a different learning task Y, each column represents a different sensitive attribute targeted by the adversary. We observe that regardless of the model architecture, for any learning task there always exists a sensitive attribute for which Δ_{Adv} and thus violates the LPP.

B.7. Across Datasets

We ran an additional experiment to demonstrate that the strict trade-off between model utility and the LPP also holds on a very different type of dataset and model. As for tabular data, together with image data, sharing feature encodings instead of raw data is often suggested as a solution to limit harmful inferences, we choose the Texas Hospital dataset (Texas Department of State Health Services, Austin, Texas, 2013) and the TabNet model architecture (Arik & Pfister, 2021) for these experiments.

Data. The Texas Hospital Discharge dataset (Texas Department of State Health Services, Austin, Texas, 2013) is a large public use data file provided by the Texas Department of State Health Services. The dataset we use consists of 5,202,376 records uniformly sampled from a pre-processed data file that contains patient records from the year 2013. We retain 18 data attributes of which 11 are categorical and 7 continuous.

Experiment Setup. In each experiment, we select one attribute as the model's learning task Y and a second attribute as the sensitive attribute S targeted by the adversary. We repeat each experiment 5 times to capture randomness of our

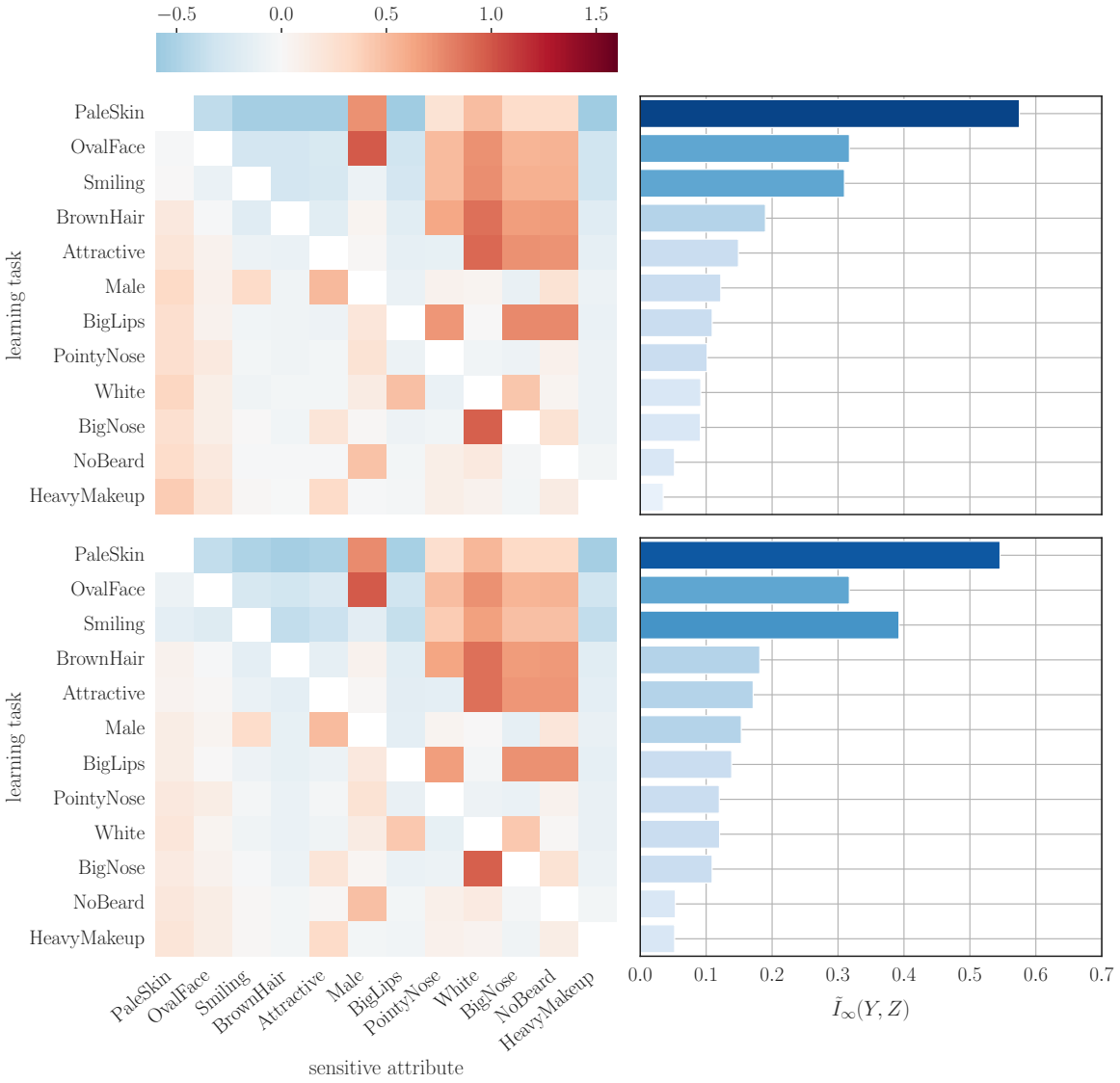

Figure 9: **The least-privilege and utility trade-off holds regardless of the model architecture.** The adversary's inference gain (*left*) always exceeds the utility gain (*right*) for at least on sensitive attribute. This hold for both a CNN256 model (*top*) and a RESNET18 model (*bottom*).

measurements for both the model and adversary, and show average results across all 5 repetitions. At the start of the experiment, we split the data into the three sets D_T, D_E, and D_A. We train a TabNet model on the train set D_T for the chosen learning task and then estimate the model's utility on the evaluation set D_E. We measure the utility of the model-generated representations as the multiplicative gain $\tilde{I}_\infty(Y; Z)$. After model training and evaluation, we train both the label-only and features adversary on the auxiliary data D_A. The features adversary is given access to a record's representation at the last encoding layer of the TabNet encoder (see (Arik & Pfister, 2021) for details of the model architecture). For a given sensitive attribute S, we estimate the adversary's gain as $\tilde{I}_\infty(S, Z \mid Y)$.

As above, the bar chart in Fig. 10 (*right*) shows the model's utility for learning task Y indicated in each row measured as $\tilde{I}_\infty(Y, Z)$. The heatmaps in Fig. 10 (*left*) show the difference between the adversary's inference gain and the model's utility $\Delta_{Adv} = \tilde{I}_\infty(S, Z \mid Y) - \tilde{I}_\infty(Y, Z)$. As on the LFWA+ dataset, for any learning task there always exists a sensitive attribute for which an adversary gains an advantage from observing a target record's feature representation and $\Delta_{\mathsf{Adv}} > 0$. This demonstrates the strict trade-off between utility and the LPP as predicted by Theorem 2.

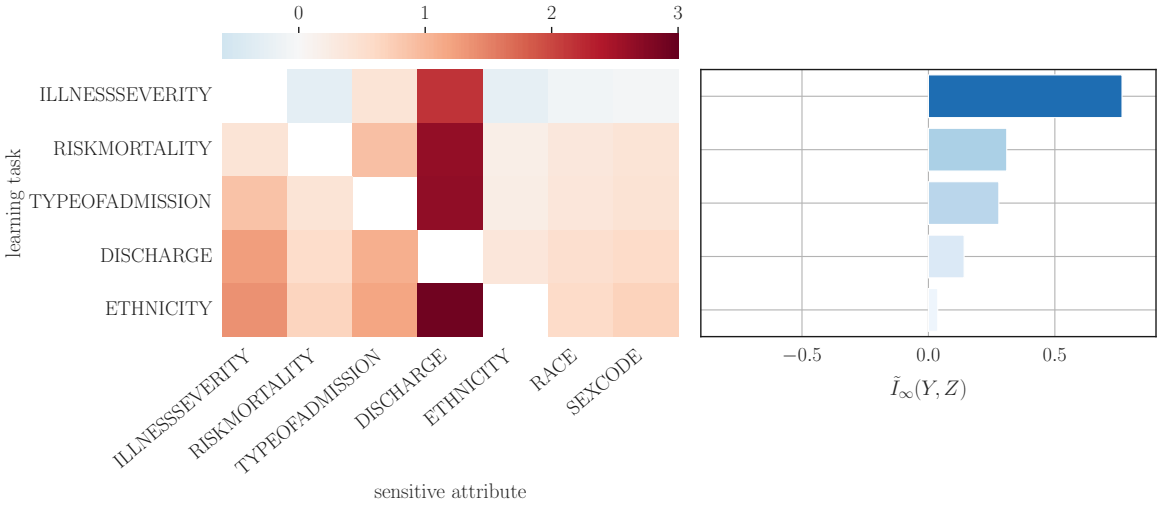

Figure 10: Attribute leakage (*left*) and model utility (*right*) for a TabNet model trained on the Texas Hospital dataset