SANITIZER: SENSITIVE INFORMATION PROTECTION FOR TASK INDEPENDENT DATA RELEASE

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose Sanitizer, a framework that protects against sensitive information leakage to facilitate task independent data release with untrusted parties. This is done in a two-step process: first, we develop a framework that encodes unstructured data into a structured representation bifurcated by sensitive and non-sensitive representation. Second, we design mechanisms that transform the sensitive features such that the leakage of sensitive information is minimal. Instead of removing sensitive information from the unstructured data, we replace the sensitive features by sampling synthetic sensitive features from the joint distribution of the sensitive features in its structured representation. Hence, using this method one can share a sanitized dataset that preserves distribution with the original dataset resulting in a good utility-privacy trade-off. We compare our technique against state-of-the-art baselines and demonstrate competitive empirical results both quantitatively and qualitatively.

1 INTRODUCTION

In the current paradigm of data driven decision making, data sharing between untrusted parties is paramount for collaboration. Few such examples include training machine learning model on crowd-sourced data, prediction and analytics over user data by cloud based service providers, etc. This data sharing has been fundamentally inhibited due to the leakage of sensitive and private information associated with individuals participating in data sharing. In this work we consider the problem of *sanitization* where sensitive information from a dataset is obfuscated in a task independent manner and then released to untrusted parties. This task independence allows the data analyst on the receiving end to ask any arbitrary query related to the dataset. Some of these queries can be invasive to the privacy of the data sharing entity, for example - the query of estimating the sensitive information itself. Therefore, the problem of protecting sensitive information while releasing data comes with an unavoidable privacy-utility trade-off. We study different constraint emerging from this problem setup and propose a framework and show its efficacy by evaluating privacy-utility trade-off under different experimental constraints.

While we focus on protecting sensitive attributes, some of the existing works in this area focusing on protecting against membership inference attacks using formal privacy guarantee obtained by differential privacy. This is a common and pertinent problem when a machine learning (ML) model trained on private data is accessible to adversary. In this work, we focus on data release instead of model release and provide a framework for the protection against sensitive information leakage instead of membership inference. The problem of context dependent information leakage is different than the traditional definitions used in the privacy community like differential privacy (Dwork et al., 2014) that provides uniform privacy guarantee. This problem is also referred as attribute inference (Jia & Gong, 2018) and has been studied quite well from both defense (Jia & Gong, 2018; Raval et al., 2019; Edwards & Storkey, 2015) and attack (Kosinski et al., 2013; Chaabane et al., 2012; Jia et al., 2017) perspectives. There are several variants of differential privacy (Huang et al., 2017; Doudalis et al., 2017; Kifer & Machanavajjhala, 2014; He et al., 2014; Liu et al., 2017) that factor in the notion of sensitive attributes in their privacy definition.

Sanitization concerns with transforming a data sample to remove *sensitive* attribute information while retaining every other information with a goal of keeping its utility high for any arbitrary downstream task. Existing methods achieve sanitization by transforming the input in one of two ways: i)

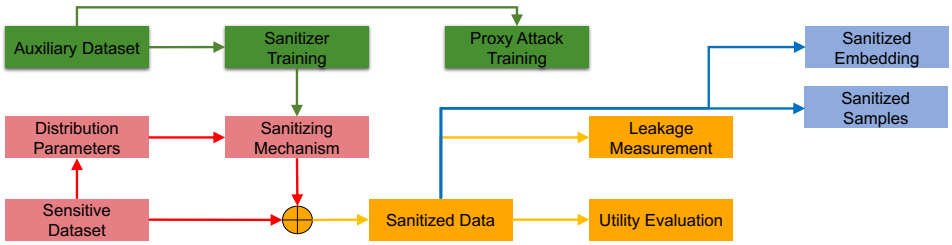


Figure 1: **Sanitizer pipeline:** First, we utilize publicly available (in green) datasets to train sanitizer models. Then we learn some particular parameters from the sensitive dataset (in pink) to design sanitizing mechanisms that is applied on the sensitive dataset to obtain sanitized data that could be potentially sensitive or not (in yellow). We evaluate the performance by measuring utility and leakage, and then the sanitized embedding and samples are released (in blue).

sanitization by information removal or ii) sanitization by sampling from data distribution. The first category of methods remove information using adversarial learning (Huang et al., 2018a; Li et al., 2020), adding noise to representations (Kotsogiannis et al., 2020; Kifer & Machanavajjhala, 2014; He et al., 2014). The second category of approaches typically learn a generative model (Jordon et al., 2018; Beaulieu-Jones et al., 2019; Xin et al., 2020; Takagi et al., 2020; Rosenblatt et al., 2021) to synthesize new samples from the same underlying data distribution but are not the exact samples from the sensitive dataset. Both sets of methods provide relatively low privacy-utility *trade-off* due to their inability to preserve the non-sensitive information while eliminating the sensitive information. The methods in the first category requires removing a lot of correlated features to achieve satisfactory sanitization. Furthermore, the methods in the second category are unable to preserve any non-sensitive information for the training datapoints since they re-sample from the distribution. At a high-level there are four key categories that differentiates our work in the existing landscape - **Output:** We release a sanitized version (embedding and sample) of each and every data sample instead of a parametric model. **Input:** While the input could be (high-dimensional and unstructured) data like images, our mechanisms apply transformation over (structured and low-dimensional) feature representation. **Threat model:** We protect against attribute inference instead of membership inference. **Mechanism:** Our mechanisms apply transformations that results a new synthetic sensitive feature which is released along with unperturbed embedding. We view this as a bridge between obfuscation based mechanisms (that aim to remove sensitive information completely) and synthetic data sampling (that re-synthesize all features).

In this work, the proposed method seeks to provide better privacy-utility trade-offs by integrating the two diverse perspectives: (i) removing local sensitive information and (ii) sampling alternatives from the attribute distribution to sanitize the sensitive datapoints. This workflow is summarized in figure 1. Central to this work is the insight that a dataset can be decomposed into a set of semantic concepts in the latent space. To this end, we design a regularized variational autoencoder (called $\alpha\beta$-VAE) which decouples the sensitive semantic concepts in latent space and samples their alternates from the corresponding label distributions. The sanitized datapoints can be regenerated from the decoder and used for downstream tasks. We also note that, for dataset specific queries, this approach can *shift* the privacy-utility frontier since the sanitized dataset can still be used to train models on the sensitive attributes while protecting the sensitive information of individual's sensitive attributes.

The contributions of this work can be summarized as follows:

- We introduce *sanitizer*, a generic framework that semantically decouples sensitive and non-sensitive representation from data. We empirically demonstrate its efficacy on multiple datasets.
- Using our framework, we design mechanisms that allow dataset release while protecting against leakage of sensitive information.
- We quantitatively and qualitatively show superior privacy-utility trade-off than existing works and also demonstrate that sanitized mechanisms make it possible to learn the distribution of sensitive attributes while preventing leakage of sensitive attribute of an individual.
- We release a dataset of sanitized representation for our proposed technique and other baseline approaches for future benchmarking to enable rigorous evaluation of attacks and defenses.

2 RELATED WORK

The work focuses on protecting leakage of sensitive information when data is released to unstrusted third parties. Correspondingly, we outline the landscape of research in privacy protection (i) and note relevant work for private data release (ii) including literature to protects against attribute inference attacks (iii). To motivate our protblem formulation and method, we also draw parallels to relevant research in fairness and semantic manipulation (iv).

Data Privacy Protection methods focus on providing privacy during model training (Abadi et al., 2016; Papernot et al., 2016), at inference (Roy & Boddeti, 2019b; Singh et al., 2020) or for release of training data (Huang et al., 2018a; Ping et al., 2017). First, methods have been developed for protecting training data when either training data is distributed over clients or computation for training the model is out-sourced. For the former, distributed learning techniques such as federated learning (Kairouz et al., 2019; Konečnỳ et al., 2016) and split learning (Gupta & Raskar, 2018; Vepakomma et al., 2018a) are used, where the clients communicate with a centralized server using weights and activations and the latter relies on homomorphic encryption (Gentry & Boneh, 2009; Brakerski et al., 2014) and secure enclaves (Zhang et al., 2020; Ferraiuolo et al., 2017). Second, private inference methods focus on collaborative setup where model is distributed across a network and communication is based on privateized intermediate activations. To achieve this, methods have sought to either decouple features of private and task attributes by minimize distance correlationVepakomma et al. (2020), mutual information using adversarial lossRoy & Boddeti (2019a); Xiao et al. (2020), or entirely prune out sensitive features (Singh et al., 2020). Finally, methods for private data release focus on protecting data before externally sharing it for arbitrary downstream use. Techniques for this can be clustered as: i) (Kotsogiannis et al., 2020; Kifer & Machanavajjhala, 2014; He et al., 2014; Li et al., 2020; Huang et al., 2018a) which add noise to the datapoint, ii) (Jordon et al., 2018; Beaulieu-Jones et al., 2019; Xin et al., 2020; Torkzadehmahani et al., 2019; Takagi et al., 2020; Rosenblatt et al., 2021) which (first learn and then) re-sample the underlying joint distribution of the datapoints for release and, iii) (Abadi et al., 2016; Papernot et al., 2016; Wu et al., 2019; Pichapati et al., 2019; McMahan et al., 2017; Nasr et al., 2018; Jia et al., 2019) which release privatized models, instead of original data.

Private Data Release methods work under two threat models, in context of machine learning: i) membership inference (Shokri et al., 2017) or ii) attribute inference (Jia & Gong, 2018). Membership inference attacks aim to identify whether a data sample was present in the training dataset of a machine learning model. Attribute inference attacks aim to infer private attributes for a public released datapoint. Methods routinely propose to obfuscate (add noise to datapoints) (Kotsogiannis et al., 2020; Kifer & Machanavajjhala, 2014; He et al., 2014; Li et al., 2020; Huang et al., 2018a) or (re)generate (learn and re-sample underlying distribution of datapoints) (Jordon et al., 2018; Beaulieu-Jones et al., 2019; Xin et al., 2020) datapoint for privatized release, largely under the threat of membership inference. Some obfuscation-based methods (Li et al., 2020; Huang et al., 2018a) have been adapted to defend against attribute inference but have provide limited utility. In particular, (Li et al., 2020) obfuscates in activation space and is not equipped to release privatized natural images. In this work, we also focus on data release under threat of attribute inference and propose sanitization techniques that bridges work in obfuscation and generative techniques to provide improved privacy-utility trade-offs when releasing image samples.

Privacy against Attribute Inference For private release of data under attribute inference threats, research has focused on both attack and defense schemes. Defense methods have proposed adding various forms of differentially private (Dwork et al., 2006a) noise and often for tabular (structured) data. We seek to extend attribute privacy to image data which presents distinct challenges. Specifically, adding DP-noise to images has been shown (Li & Clifton, 2021; Singh et al., 2020) to provide poor privacy-utility trade-offs due to lack of structure in input modality and high sensitivity. This is because DP provides uniform protection to all sensitive and non-sensitive attributes, significantly hampering utility of transformed input. This work alleviates the same through a sanitization scheme that seeks to only remove (by suppressing, obfuscating or re-sampling) sensitive attribute information in latent space and then (re)generating the transformed datapoint for privatized release.

Fairness and Semantic Manipulation Our *problem formulation* for sanitization is similar to learning fair and unbiased representations (Zemel et al., 2013; Sarhan et al., 2020; Creager et al., 2019; Zhang et al., 2018; Mehrabi et al., 2021) and the formalism of sensitive information also aligns to

notion of "nuisance variable" in domain adaptation (Louizos et al., 2015). However, in constrast to methods that seek to learn *latent representations* that are invariant to nusiance variable, sanitizer aims to prevent leakage from released *input* data. Hence, in this work, the sanitization problem is stronger setup and cannot directly utilize ideas designed for fairness. Our *proposed method* for sanitization is similar to work in semantic manipulation (Shen & Liu, 2017; Chen et al., 2019). We note that while this has resulted in some relevant work in privacy, there are is primarily restricted to faces and methods are designed with objective to support a **specific utility attribute** like identity (Othman & Ross, 2014; Chhabra et al., 2018; Mirjalili et al., 2018; Morales et al., 2020; Mirjalili et al., 2020; Li et al., 2021; Mirjalili et al., 2019; Othman & Ross, 2014; Raynal et al., 2020). This is very different from sanitizer which is **agnostic of target utility and only depends upon sensitive attribute**. We also note some recent work on sensitive attribute which uses uncertainty based techniques Wang et al. (2021) and Martinsson et al. (2020) to obfuscate sensitive attribute in latent space using adversarial training but hence, unlike sanitizer, generates highly unrealistic images due to high uncertainty.

3 Problem Formulation

Terminology: Consider a data holder A with access to a dataset $D = \{\mathbf{X}, \mathbf{Y}\}$ with N datapoints. Let $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ represent a pair of input sample ($\mathbf{x}$) and corresponding *set* of labels that describe distinct attributes of the input. For instance, given $\mathbf{x}$ is a face image of an individual, the *set* $\mathbf{y}$ may include the age, keypoints and ethnicity for that individual. From standpoint of the data holder, certain attributes in the label set $\mathbf{y}$ may represent sensitive information (called $\mathbf{y}_S$) while others are non-sensitive ($\mathbf{y}_{NS}$) such that $\mathbf{y} = \{\mathbf{y}_S \cup \mathbf{y}_{NS}\}$. This sensitive attributes often represents private information which A does not want to share. Hence, if the dataset D is to be released for external modeling, a key challenge is to protect privacy of the sensitive information while preserving utility of the non-sensitive information. Enabling this private data-release while ensuring optimal privacy-utility trade-offs is the central focus of this work. Our proposed method achieves this through two key phases: i) *semantic decoupling* in the latent space of $\mathbf{x}$, the representations of sensitive ($\mathbf{z}_S$ for $\mathbf{y}_S$) and non-sensitive attributes ($\mathbf{z}_{NS}$ for $\mathbf{y}_{NS}$) using a regularized β-VAE called $\alpha\beta$-VAE (section 4.1) and ii) *sanitizing* the sensitive representation $\tilde{\mathbf{z}}_S = f(\mathbf{z}_S)$ using a mechanism (f) that obfuscates the sensitive attributes ($\mathbf{y}_S$) to generate $\tilde{\mathbf{y}}_S$ (section 4.2). Finally, we concatenate $(\mathbf{z}_{NS}||\tilde{\mathbf{z}}_S)$ and obtain the sanitized embedding $\tilde{\mathbf{z}}$ and sanitized sample $\tilde{\mathbf{x}}$ which can be released for downstream tasks.

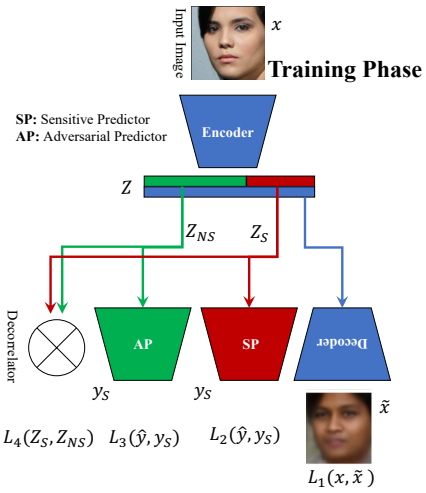


Figure 2: Training scheme for the proposed $\alpha\beta$-VAE. x is the input sample, $\hat{y}$ and are the predicted and ground-truth sensitive attribute(s) respectively. and are the latent representations for sensitive and non-sensitive information. The four objectives used are detailed in Section 3

Threat Model: We assume the *untrusted* data-receiver B can act as an adversary by obtaining access to a subset of the dataset $\tilde{D} = \{\tilde{\mathbf{X}}, \{\mathbf{Y}_S, \mathbf{Y}_{NS}\}\}$. Note that this leaked dataset contains a mapping from sanitized inputs $\tilde{\mathbf{X}}$ to sensitive labels $\mathbf{Y}_S$ and not $\tilde{\mathbf{Y}}_S$. Thus, the attacker can train an ML model on the leaked dataset and then infer sensitive attributes. Since the goal of this work is to preserve utility for dataset specific queries, we also assume that the attacker knows a prior on the sensitive attribute distribution ($p(\mathbf{y}_S, \mathbf{y}_{NS})$).

Defining information leakage: Information leakage for sanitization has been typically defined through information theoretic terms (Makhdoumi & Fawaz, 2013; Sankar et al., 2010), however, estimating these terms requires estimation of probability distributions which is not tractable for higher dimensions therefore we use a data driven approach to quantify leakage. The goal of our framework is to minimize the change in belief before (prior $p(\mathbf{y}_S)$) and after (posterior $p(\mathbf{y}_S|\tilde{\mathbf{x}})$) observing the sanitized data ($\tilde{\mathbf{x}}$) as measured by the prediction performance over the sensitive attribute over the test sample. This notion has been formalized in anonymization literature as information theoretic privacy (Rebollo-Monedero et al., 2009) and bayes-optimal privacy (Machanavajjhala et al., 2007).

Desiderata: We want the following three properties from our sanitizer framework: **P1**) $q(\mathbf{z}_S, \mathbf{z}_{NS}|\mathbf{x}) = q(\mathbf{z}_S|x)q(\mathbf{z}_{NS}|x)$ (Independence among the sensitive and non-sensitive latent representation), **P2**) $p(\mathbf{x}|\tilde{\mathbf{z}})$ is maximized, and **P3**) $q(\tilde{\mathbf{z}}_S)$ is similar to $q(\mathbf{z})$. The first property would enable *semantic decoupling*, while the second property allows high utility and the third property ensures the distribution of sensitive information is preserved. We utilize variational auto-encoders (VAEs) as our building block to attain these three properties and discuss it below.

VAE: Given a dataset $\mathbf{X}$, Variational autoencoders (VAE) (Kingma & Welling, 2013; Rezende et al., 2014), are used to model the distribution of samples $p(x) \forall \mathbf{x} \in \mathbf{X}$ by learning parameters ϕ of approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and θ for the likelihood $p_\theta(\mathbf{x}|\mathbf{z})$. Higgins et al. (2016) improve the disentanglement between the components of samples from $q_\phi(\mathbf{z}|\mathbf{x})$ by regularizing the KL divergence between the prior $p_\theta(\mathbf{z})$ and approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$. To obtain the property *P2*, we need a high degree of disentanglement between $\mathbf{z}_S$ and $\mathbf{z}_{NS}$. To improve disentanglement, existing works such as Kim & Mnih (2018) and Chen et al. (2018) regularize total correlation of $q(z)$ that is measured as $\mathsf{KL}(q(z)||\Pi_{i=1}^{m} q(z_j))$ where KL refers to KL divergence and m is the total number of components of z. Since none of our desired properties require disentanglement between every component of z but rather only the disentanglement between $\mathbf{z}_S$ and $\mathbf{z}_{NS}$ (property P1), we propose a new regularized $\alpha\beta$-VAE.

A key characteristics of VAE is that the decoupled latent representations are unordered. This means that there is no explicit control on which dimensions encode which semantic attribute. This is a challenge for our work that ideally requires that representations encoding the sensitive attributes be contiguous to avoid information leakage. To realise this, we reformulate the objective with a *aligner* $g_u(\cdot,)$ parameterized by u that is trained to estimate $\mathbf{y}_S$ from $\mathbf{z}_S$, the intuition is that backproping *aligner*'s gradient from $q(\phi)$ will force to maximize relevant information between $p(\mathbf{y}_S|\mathbf{x})$ and $q(\mathbf{z}_S|\mathbf{x})$. Since all latents are known to be correlated with each other to a certain extent, we need to prevent information leakage about $\mathbf{y}_S$ in $q(\mathbf{z}_{NS})$. Instead of regularizing total correlation between each dimension, we propose to regularize correlation between $q(\mathbf{z}_S)$ and $q(\mathbf{z}_{NS})$. We use distance correlation (Székely et al., 2007; Székely & Rizzo, 2013; Székely et al., 2014; Vepakomma et al., 2018b) as the target objective to be minimized. Since directly estimating probability density is intractable for high dimensional representations, various measures such as HSIC (Gretton et al., 2005; Sejdinovic et al., 2013), MMD (Borgwardt et al., 2006) and distance correlation (Roy & Boddeti, 2019b; Sadeghi et al., 2019; Huang et al., 2018b; Li et al., 2019) are used. In particular, we use distance correlation ($dcorr$) since it enables enables measuring nonlinear correlations between samples from random variables of arbitrary dimensions ($\mathbf{z}_S$ and $\mathbf{z}_{NS}$ can have different dimensionality), allows for efficient gradient computation and does not require any kernel selection or parameter tuning unlike HSIC and MMD. However, we do note that $dcorr$ is measured between samples and hence larger sample size is desirable for the unbiased sample statistic to represent the population notion of distance correlation. The final objective can be summarized as:

$$L_1(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{1}$$

$$L_2(\phi, u) = \ell_1(g_u(\mathbf{z}_i \sim q_\phi(\mathbf{x})|_{i \leq k}), \mathbf{y}_S) \tag{2}$$

$$L_3(\phi) = dcorr(\mathbf{z}_i \sim q_\phi(\mathbf{x})|_{i \leq k}, \mathbf{z}_i \sim q_\phi(\mathbf{x})|_{k < i \leq m}) \tag{3}$$

Here L_1 is the β-VAE (Higgins et al., 2016) formulation of VAE's lower bound. L_2 is the objective for training the parameters of the aligner model and L_3 is the training objective for minimizing correlation between $\mathbf{z}_S$ and $\mathbf{z}_{NS}$. Optimizing L_2 and L_3 jointly help us in achieving the desired property P1 and P2. This can be further regularized by an adversarial prediction network $h_v(\cdot,)$ parameterized with v which constrains $\mathbf{z}_{NS}$ to diverge from $\mathbf{y}_S$. However, as shown later in the experiments, this adversarial regularizer does not significantly impact the leakage.

$$L_4(\phi, v) = \ell_2(h_v(\mathbf{z}_i \sim q_\phi(\mathbf{x})|_{k < i \leq m}), \mathbf{y}_S) \tag{4}$$

The loss functions l_1, l_2 can be cross-entropy or l_p-norm (often $p = 2$) depending upon $\mathbf{y}_S$.

Finally, the parameters ϕ, θ, u are trained jointly as specified in the equations 1-4 under a simultaneous optimization (Mescheder et al., 2017) framework. The overall training objective is given as:

$$\arg\min_{\theta, \phi, u} \alpha_1 L_1(\theta, \phi) + \alpha_2 L_2(\phi, u) + \alpha_3 L_3(\phi) - \alpha_4 L_4(\phi, v) \tag{5}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are scalar hyper-parameters. This completes our description of the proposed $\alpha - \beta$VAE. Now we discuss the mechanisms for designing f.

Method	Fairface ↑	CelebA ↑	UTKFace ↑
TIPRDC (Li et al., 2020)	0.441	0.465	0.453
Adversarial (Huang et al., 2018a)	0.447	0.442	0.450
Noise	0.438	0.422	0.435
Adversarial Noise (Huang et al., 2018a)*	0.432	0.422	0.420
Ours	**0.476**	**0.483**	**0.476**

Table 1: **Privacy-Utility comparison:** We report *normalized hypervolume* scores (higher is better) to compare the privacy-utility trade-offs for our sanitizer framework against baseline. '*' we adapt the method in original paper for this task with implementation details in section 4.2. Proposed method outperforms all baselines in all experiments.

3.1 SANITIZING THE SENSITIVE SPLIT

Here, the key idea is to apply function $\tilde{\mathbf{z}}_S = f(\mathbf{z}_S)$ where $\mathbf{z}_S, \mathbf{z}_{NS} \sim q_\phi(\mathbf{x})$. We consider four alternate mechanisms for transforming sensitive features $\mathbf{z}_S$ which we describe below. We compare these schemes, qualitatively and quantitatively, in section 5.

a) Suppression: In this scheme, we explicitly remove the sensitive information by replacing $\mathbf{z}_S$ with a zero vector (i.e. $\tilde{\mathbf{z}}_S$ is a zero vector). Since the prior on $p(z) = \mathcal{N}(0, I)$, setting up $\tilde{\mathbf{z}}_S = 0$ still results in plausible reconstructions by the likelihood parameter θ.

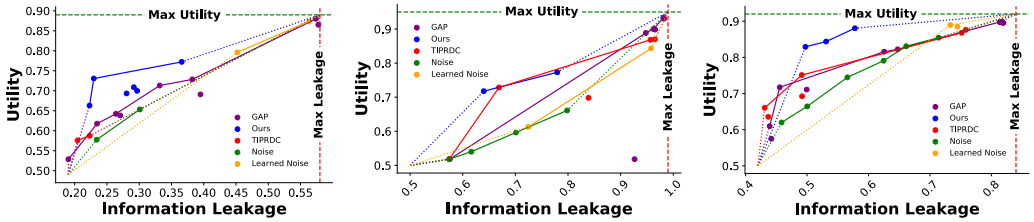
b) Obfuscation: Inspired by noise based mechanisms used in differential privacy (Dwork et al., 2006b), we add laplace noise to $\mathbf{z}_S$, i.e. $f(\mathbf{z}_S) = \mathbf{z}_S + r$ where elements of r are sampled iid from laplace distribution. This is represented as $r \sim Lap(\frac{\Delta}{\epsilon})$ here Δ is the sensitivity and ϵ is the privacy budget. We ensure the norm of $\mathbf{z}_S$ is always clipped to a prescribed value prior to the release. This ensures the global sensitivity is always known before hand with respect to the architecture. This holds as long as the model and the sanitizing pipeline remains with the model owner and only the sanitized samples produced by the system are released to the user. Note that the noise is added dimension wise and is precisely known as the multidimensional Laplace mechanism (Dwork et al., 2006b). A caveat however is that, since the noise is directly proportional to the dimensionality of the $\mathbf{z}_S$ i.e. k, the amount of noise scales up quickly with the increase in k. Since linear transformation in the latent space of VAE is amenable to reconstruction quality (Kingma & Welling, 2013), our obfuscation mechanism results in realistic samples from the data distribution. Note that this only gives privacy guarantee for releasing $\mathbf{z}_S$, however, $\mathbf{z}_{NS}$ can still carry sensitive information.

c) Generalization: In this mechanism, we obtain the k clusters of $\mathbf{Z}_S$ and then obtain $\tilde{\mathbf{z}}_S$ as a vector of sensitive features that is uniformly chosen as one of the cluster mean. While this approach can attain good reconstruction quality, the reconstructed dataset would lack diversity that could arise from intra-cluster features of $\mathbf{z}_S$ since we only use centroids.

d) Joint sampling: For this sampling mechanism, we sample sensitive features for each data point from the joint distribution of the sensitive features $p(\mathbf{z}_S)$. While the prior $p_\theta(\mathbf{z})$ is an isotropic gaussian, the approximate posterior $q(\mathbf{z})$ takes on a joint distribution with a non-diagonal covariance matrix. Sampling from the prior $p_\theta(\mathbf{z})$ could result in objectives related to fairness but the goal of this work is to preserve prior statistics of a sensitive dataset to ensure good performance on dataset specific query. Therefore, we learn this joint distribution using a gaussian mixture model (GMM). The main reason to use GMM is efficient sampling and its capability to learn small dimensional datasets ($\mathbf{Z}_S$ in this case). Therefore in this case, the function f is parameterized by π_c, μ_c, Σ_c for every component c. We obtain $\Pr[\tilde{\mathbf{z}}_S] = \sum_c \pi_c \mathcal{N}(\tilde{\mathbf{z}}_S | \mu_c, \Sigma_c))$ by sampling from the GMM model. With the newly obtained $\tilde{\mathbf{z}}_S$, we compute $\tilde{\mathbf{y}}_S = g_u(\tilde{\mathbf{z}}_S)$ and $\tilde{\mathbf{x}} \sim p_\theta(\tilde{\mathbf{z}}))$. Each sample $x \in X$ is obtained independently and finally the sanitized dataset $\tilde{D} = \{\tilde{X}, \{\tilde{\mathbf{Y}}_S, \mathbf{Y}_{NS}\}\}$ is shared with B. We note that, in contrast to prior work (Huang et al., 2018a; Li et al., 2020; Liu et al., 2017), a key benefit of this strategy is in providing the flexibility to protect the sensitive information for individual data points while learning a classifier for the sensitive attribute using the privatized dataset.

4 EXPERIMENTS

In this section, we first introduce the datasets, baselines, implementation details, evaluations and metrics. Finally, we compare our method with the state-of-the-art and baselines methods through quantitative and qualitative evaluation. While the proposed techniques are task independent, we choose utility attribute for every dataset.



(a) FairFace (b) CelebA (c) UTKFace

Figure 3: **Privacy-utility trade-off evaluation on different datasets:** We plot sensitive information leakage as a proxy for privacy and one of the task attribute as a measure of utility for the sanitized dataset. Each point in this plot corresponds to training a sanitizer model and then evaluating its performance by training adversary model and utility model on the sanitized dataset. Sanitizer performs better than all existing methods on all three datasets. We report the normalized hyper-volume in table 1 as a metric to compare the privacy-utility trade-off for different methods.

4.1 DATASETS

UTKFace (Zhang & Qi, 2017) consists of 20,000 face images. We use the cropped and aligned version of the dataset and generate a random split of $90\% - 10\%$, training and testing data. The dataset has ethnicity, gender, and age as categorical labels. For our experiments, we keep the sensitive attribute as ethnicity which has 5 unique labels and due to class imbalance, best possible performance without access to the image is 44%. We use gender as the utility attribute.
CelebA (Liu et al., 2018) is a large scale dataset of 202,599 celebrity face images 10,177 unique identities, each with 40 binary attribute annotations. For our experiments, we define gender as the sensitive attribute. We use *mouth open* as an attribute for utility evaluation.
FairFace (Kärkkäinen & Joo, 2019) dataset consists of 108,501 images, with three different attributes race, gender, and age. The dataset was curated to overcome the class imbalance in ethnicity present in the majority of the existing Faces datasets. We use ethnicity as sensitive attribute and gender as utility attribute for our experiments.
Dataset and benchmark release: To encourage further work in private data release, we create a benchmark dataset of 1-million sanitized images by applying baselines on existing datasets and baselines. This will enable rigorous evaluation of different mechanisms and their privacy-utility trade-off. The benchmark will serve as a continuously improving evaluation pipeline for researchers to study both attack or defense techniques. We plan to release the benchmark and datasets after receiving the feedback from reviewers.
All of our experiments are implemented using PyTorch (Paszke et al., 2019) and conducted using NVIDIA 1080 Ti GPUs. We use Adam optimizer (Kingma & Ba, 2014) for training all of the neural networks. We perform all of our experiments in three phases similar to the pipeline described in the figure 1. In the first phase we train the sanitizer and similar analogous networks for our baselines. Second, we sanitize the target dataset based on the mechanism. Finally, we train two separate ML models on the sanitized dataset for evaluating utility and information leakage. We plan to release all of our experimental code, datasets, models and experimental configurations post review process as part of the benchmark.

4.2 BASELINES

We compare our proposed method with state-of-the-art visual sanitization techniques GAP (Huang et al., 2018a) and TIPRDC (Li et al., 2020), and introduce new baselines for exhaustive comparison. The key baselines are described next: i) *GAP (Huang et al., 2018a):* is trained adversarially to maximize loss for a proxy adversary trying to infer sensitive attribute on the sanitized images. We tune the hyperparameter λ for controlling reconstruction quality vs the proxy adversary performance. We modify the architecture proposed in the original paper to improve their results for the datasets used in our paper. ii) *Learned Noise:* is built upon the TCNND architecture described in GAP (Huang et al., 2018a) where small dimensional noise is fed to a decoder that generates a vector of the same dimensionality of the raw image and sanitized image is obtained by adding up the two. iii) *TIPRDC (Li et al., 2020):* is used as a baseline without any modification, here again we vary the parameter λ to obtain trade-off. While we make quantitative comparison with TIPRDC, it is not possible to compare it in our qualitative results since the sanitized dataset released by TIPRDC is activations of an intermediate layer of a neural network and not images. iv) *Noise:*

	UTKFace		CelebA	
	Utility	Leakage	Utility	Leakage
Uniform Noise	0.667	0.501	0.576	0.712
Adversarial	0.615	0.499	0.723	0.686
Adversarial Noise	0.801	0.695	0.746	**0.663**
Sanitizer	**0.86**	**0.474**	**0.9022**	0.6955

Table 2: **Classification Accuracy Score (CAS) evaluation**: We train classifier on privatized data samples and evaluate them on non-privatized samples.

	UTKFace		CelebA	
	Utility	Leakage	Utility	Leakage
Suppression	0.208	0.498	0.7042	0.7177
Generalization	0.387	0.672	**0.8764**	0.821
Obfuscation	0.208	0.491	0.62	0.7129
Sampling	**0.521**	**0.474**	0.817	**0.6955**

Table 3: **CAS on sensitive attribute estimation as utility**. Note that this setup is not possible for the baselines hence we only evaluate our proposed four mechanisms.

baseline sanitizes the image by adding gaussian noise in the pixel space of raw image directly. We keep the mean 0 and vary standard deviation σ to obtain trade-off.

4.3 Evaluation and Metrics

We report performance using **Privacy-Utility trade-off** which compares the capability of an adversary to correctly infer sensitive information from the sanitized representation which is concurrently used by a user to infer task information. For this analysis, we simulate a *worst case* adversary that has the chance to dynamically adapt to the privatization scheme, which is modeled by finetuning the adversary on a *privatized* validation set and then evaluating of a disjoint test set.

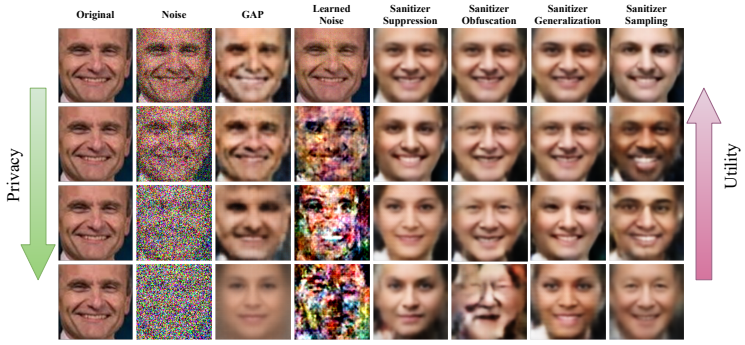


Figure 4: **Qualitative demonstration of privacy-utility trade-off:** We tune trade-off parameter for the sampling mode of *sanitizer* and baseline techniques on the UTKFace dataset. This results in different reconstruction quality and sensitive information leakage for the generated images. The privacy axis here is measured based on adversary performance on the sensitive attribute (ethnicity) prediction over the test set.

Experiment E1: We evaluate privacy and utility trade-off using accuracy of adversary and user, respectively, and represent the trade-off using a normalized hyper-volumue (HV) (Ishibuchi et al., 2018) (inspired by (Roy & Boddeti, 2019b)).

Experiment E2: To measure the usefulness and sample quality, we experiment with training sanitizer and other baselines on privatized images and test it on non-privatized images. A good result on the test set would indicate a reasonable amount of domain transfer. This is referred as Classification Accuracy Score(CAS) in the generative model community (Ravuri & Vinyals, 2019). Note that it is not possible to include TIPRDC since their output is constrained to embedding sharing.

Experiment E3: We perform a similar experiment as E2 but this time the attacker tries to estimate sensitive information of sanitized data while a user computes CAS on sensitive dataset during the test phase. This setup is motivate by a practical scenario where data is crowdsourced between untrusted users and then deployed on trusted devices.

5 Results

Quantitative Results: We compare privacy-utility trade-off on all three datasets with our baselines by varying the corresponding trade-off parameter and plotting it in figure 3. Our *sanitizer* framework obtains a better privacy-utility trade-off as measured by the normalized hyper-volume. For evaluating the plausibility of building ML models using the sanitized dataset, we train a standard ResNet-34 (He et al., 2015) on sanitized datasets obtained by our sanitizer framework and baselines. Then we compare its test set performance on the unsanitized version of the dataset.

Qualitative Results: We obtain different sanitized samples for our method and baselines by adjusting trade-off parameters. We visualize the sanitized images in the figure 4. There are two key inferences from this result - i) our sanitizer protects image semantics while obfuscating the sensitive information (ethnicity); this is made possible by our mechanism's capability to transform semantics associated with sensitive attribute instead of obfuscating that inadvertently disrupts other semantic information not associated with the sensitive information such as emotion, pose etc. ii) adversarial learning based approaches do not offer any privacy protection in the high utility regime as the sensitive information (ethnicity) gets disclosed to a reasonable extent.

6 DISCUSSION

Here, we quantitatively analyse effect of the regularized introduced in $\alpha\beta$-VAE and discuss some architectural limitations. We discuss connections with existing privacy frameworks in appendix.

i) Impact of regularizers used: We perform ablation on each of the component described in the architecture in figure. 2. We measure the change in the sensitive information leakage by comparing performance with and without each component in the loss function. This can be interpreted as setting up $\alpha_i = 0$ for the i'th component during the training phase of VAE.
First we measure sensitive information leakage with all components of our regularizer connected, the sensitive attribute prediction accuracy is 22.16%.
—*Distance correlation ($dcorr$):* Removing distance correlation results in the sensitive information leakage as 29.7% depicting that removing distance correlation increases the information leakage of the sensitive attribute ($\mathbf{y}_S$) in $\mathbf{z}_{NS}$.
—*Adversarial predictor (h):* Similar to ablating the distance correlation loss function, we see increase in the information leakage of sensitive attribute in the $\mathbf{z}_{NS}$ as the accuracy of sensitive attribute prediction goes to 28.04%.
—*Sensitive predictor (g):* Ablating this component results results in marginal change in performance (sensitive attribute prediction accuracy increase to 23.5%) indicating its contribution might be lesser than the other two components for protecting against leakage.

ii) Architectural Limitations: The key goal of this work is to introduce systematic framework and mechanisms for sanitizing data that could be useful for as many downstream tasks as possible under the privacy-utility trade-off. While we seek to achieve that, here, we note two key limitations of the presented results, emerging from the generative modeling framework: i) *input sample size* - This stems from need of sufficient datapoints of to learn the semantic representation. We note that designing VAEs that can capture the distribution with few samples is active area of research in few-shot learning which will improve impact of our results but is orthogonal to scope of this work. ii) *output sample quality* - This can be improved using hierarchical VAEs and we consider this as part of future work. We also mention that while GANs are known for high fidelity image synthesis, we specifically use VAEs due to their ability to disentangle semantics in latent space, which is key to the primary objective of this work in enabling private data release.

6.1 CONCLUSION

In this work we presented sanitizer: a framework for protecting against sensitive information leakage for data release. We achieve this objective by a two step process - i) creating semantic splits and applying mechanisms that aim to preserve the underlying data distribution. While our approach demonstrates a good privacy-utility trade-off, it is possible to further decrease the sensitive information leakage in the sanitized version by improving regularization and the training method used in this paper. We presented four different mechanisms that are possible under our framework and we believe this can open up an avenue for the future researchers to develop new mechanisms that leverage the benefits semantic splits proposed in our paper. One possible future direction of this work can be to extend it to other tasks where unstructured data is involved such as speech, NLP and time series data. Finally, our framework takes first steps in the direction of bridging obfuscation and generative modeling based methods for private data release.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. (leveled) fully homomorphic encryption without bootstrapping. *ACM Transactions on Computation Theory (TOCT)*, 6(3):1–36, 2014.

Abdelberi Chaabane, Gergely Acs, Mohamed Ali Kaafar, et al. You are what you like! information leakage through users' interests. In *Proceedings of the 19th annual network & distributed system security symposium (NDSS)*. Citeseer, 2012.

Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

Ying-Cong Chen, Xiaohui Shen, Zhe Lin, Xin Lu, I Pao, Jiaya Jia, et al. Semantic component decomposition for face attribute manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9859–9867, 2019.

Saheb Chhabra, Richa Singh, Mayank Vatsa, and Gaurav Gupta. Anonymizing k-facial attributes via adversarial perturbations. *arXiv preprint arXiv:1805.09380*, 2018.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR, 2019.

Stelios Doudalis, Ios Kotsogiannis, Samuel Haney, Ashwin Machanavajjhala, and Sharad Mehrotra. One-sided differential privacy. *arXiv preprint arXiv:1712.05888*, 2017.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography*, Lecture Notes in Computer Science, pp. 265–284, Berlin, Heidelberg, 2006b. Springer. ISBN 978-3-540-32732-5. doi: 10.1007/11681878_14.

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

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Andrew Ferraiuolo, Andrew Baumann, Chris Hawblitzel, and Bryan Parno. Komodo: Using verification to disentangle secure-enclave hardware from software. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pp. 287–305, 2017.

Craig Gentry and Dan Boneh. *A fully homomorphic encryption scheme*, volume 20. Stanford university Stanford, 2009.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. *CoRR*, abs/1810.06060, 2018. URL http://arxiv.org/abs/1810.06060.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.

Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: Tuning privacy-utility tradeoffs using policies. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pp. 1447–1458, 2014.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. Context-Aware Generative Adversarial Privacy. *Entropy*, 19(12):656, December 2017. ISSN 1099-4300. doi: 10.3390/e19120656. URL http://arxiv.org/abs/1710.09549. arXiv: 1710.09549.

Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. Generative adversarial privacy. *CoRR*, abs/1807.05306, 2018a. URL http://arxiv.org/abs/1807.05306.

Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. Generative adversarial privacy. *arXiv preprint arXiv:1807.05306*, 2018b.

Hisao Ishibuchi, Ryo Imada, Yu Setoguchi, and Yusuke Nojima. How to specify a reference point in hypervolume calculation for fair performance comparison. *Evolutionary computation*, 26(3): 411–440, 2018.

Jinyuan Jia and Neil Zhenqiang Gong. Attriguard: A practical defense against attribute inference attacks via adversarial machine learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pp. 513–529, 2018.

Jinyuan Jia, Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. Attriinfer: Inferring user attributes in online social networks using markov random fields. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1561–1569, 2017.

Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 259–274, 2019.

James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2018.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)*, 39(1):1–36, 2014.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110 (15):5802–5805, 2013.

I. Kotsogiannis, S. Doudalis, S. Haney, A. Machanavajjhala, and S. Mehrotra. One-sided Differential Privacy. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 493–504, April 2020. doi: 10.1109/ICDE48307.2020.00049. ISSN: 2375-026X.

Ang Li, Jiayi Guo, Huanrui Yang, and Yiran Chen. Deepobfuscator: Adversarial training framework for privacy-preserving image classification. *arXiv preprint arXiv:1909.04126*, 2019.

Ang Li, Yixiao Duan, Huanrui Yang, Yiran Chen, and Jianlei Yang. Tiprdc: Task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and; Data Mining*, KDD '20, pp. 824–832, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403125. URL https://doi.org/10.1145/3394486.3403125.

Jingzhi Li, Lutong Han, Hua Zhang, Xiaoguang Han, Jingguo Ge, and Xiaochun Cao. Learning disentangled representations for identity preserving surveillance face camouflage. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 9748–9755. IEEE, 2021.

Tao Li and Chris Clifton. Differentially private imaging via latent space manipulation. *arXiv preprint arXiv:2103.05472*, 2021.

Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Deeprotect: Enabling inference-based access control on mobile sensing applications. *CoRR*, abs/1702.06159, 2017. URL http://arxiv.org/abs/1702.06159.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

Ali Makhdoumi and Nadia Fawaz. Privacy-utility tradeoff under statistical uncertainty. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1627–1634. IEEE, 2013.

John Martinsson, Edvin Listo Zec, Daniel Gillblad, and Olof Mogren. Adversarial representation learning for synthetic replacement of private attributes. *arXiv preprint arXiv:2006.08039*, 2020.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. *arXiv preprint arXiv:1705.10461*, 2017.

Vahid Mirjalili, Sebastian Raschka, Anoop Namboodiri, and Arun Ross. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In *2018 International Conference on Biometrics (ICB)*, pp. 82–89. IEEE, 2018.

Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. *IEEE Access*, 7:99735–99745, 2019.

Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Privacynet: semi-adversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing*, 29:9400–9412, 2020.

Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2158–2164, 2020.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 634–646, 2018.

Asem Othman and Arun Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. In *European Conference on Computer Vision*, pp. 682–696. Springer, 2014.

Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

Haoyue Ping, Julia Stoyanovich, and Bill Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pp. 1–5, 2017.

Nisarg Raval, Ashwin Machanavajjhala, and Jerry Pan. Olympus: Sensor privacy through utility aware obfuscation. *Proceedings on Privacy Enhancing Technologies*, 2019(1):5–25, 2019.

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *arXiv preprint arXiv:1905.10887*, 2019.

Mathilde Raynal, Radhakrishna Achanta, and Mathias Humbert. Image obfuscation for privacy-preserving machine learning. *arXiv preprint arXiv:2010.10139*, 2020.

David Rebollo-Monedero, Jordi Forne, and Josep Domingo-Ferrer. From t-closeness-like privacy to postrandomization via information theory. *IEEE Transactions on Knowledge and Data Engineering*, 22(11):1623–1636, 2009.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Lucas Rosenblatt, Xiaoyan Liu, Samira Pouyanfar, Eduardo de Leon, Anuj Desai, and Joshua Allen. Differentially private synthetic data: Applied evaluations and enhancements, 2021. URL https://openreview.net/forum?id=ABZSAe9gNeg.

Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019a.

Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. 2019b.

Bashir Sadeghi, Runyi Yu, and Vishnu Naresh Boddeti. On the global optima of kernelized adversarial representation learning. *CoRR*, abs/1910.07423, 2019. URL `http://arxiv.org/abs/1910.07423`.

Lalitha Sankar, S Raj Rajagopalan, and H Vincent Poor. An information-theoretic approach to privacy. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1220–1227. IEEE, 2010.

Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *European Conference on Computer Vision*, pp. 746–761. Springer, 2020.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pp. 2263–2291, 2013.

Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4030–4038, 2017.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017.

Abhishek Singh, Ayush Chopra, Vivek Sharma, Ethan Garza, Emily Zhang, Praneeth Vepakomma, and Ramesh Raskar. Disco: Dynamic and invariant sensitive channel obfuscation for deep neural networks. *arXiv preprint arXiv:2012.11025*, 2020.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

Gábor J Székely and Maria L Rizzo. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213, 2013.

Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

Gábor J Székely, Maria L Rizzo, et al. Partial distance correlation with methods for dissimilarities. *Annals of Statistics*, 42(6):2382–2412, 2014.

Shun Takagi, Tsubasa Takahashi, Yang Cao, and Masatoshi Yoshikawa. P3gm: Private high-dimensional data release via privacy preserving phased generative model. *arXiv preprint arXiv:2006.12101*, 2020.

Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564*, 2018a.

Praneeth Vepakomma, Chetan Tonde, Ahmed Elgammal, et al. Supervised dimensionality reduction via distance correlation maximization. *Electronic Journal of Statistics*, 12(1):960–984, 2018b.

Praneeth Vepakomma, Abhishek Singh, Otkrist Gupta, and Ramesh Raskar. Nopeek: Information leakage reduction to share activations in distributed deep learning, 2020.

Hui-Po Wang, Tribhuvanesh Orekondy, and Mario Fritz. Infoscrub: Towards attribute privacy by targeted obfuscation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3281–3289, 2021.

Bingzhe Wu, Shiwan Zhao, Guangyu Sun, Xiaolu Zhang, Zhong Su, Caihong Zeng, and Zhihong Liu. P3sgd: Patient privacy preserving sgd for regularizing deep cnns in pathological image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2099–2108, 2019.

Taihong Xiao, Yi-Hsuan Tsai, Kihyuk Sohn, Manmohan Chandraker, and Ming-Hsuan Yang. Adversarial learning of privacy-preserving and task-oriented representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12434–12441, 2020.

Bangzhou Xin, Wei Yang, Yangyang Geng, Sheng Chen, Shaowei Wang, and Liusheng Huang. Private fl-gan: Differential privacy synthetic data generation based on federated learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2927–2931. IEEE, 2020.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Fan Zhang, Ziyuan Liang, Cong Zuo, Jun Shao, Jianting Ning, Jun Sun, Joseph K Liu, and Yibao Bao. hpress: A hardware-enhanced proxy re-encryption scheme using secure enclave. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2020.

Song-Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

A Appendix

A.1 Connection with K-anonymity based mechanisms

The key idea in k-anonymity (Sweeney, 2002) is to separate columns of a database into two parts - non-sensitive columns (also referred as quasi-identifiers) and sensitive columns. Then a transformation is applied to the database such that any given record in the non-sensitive column should be indistinguishable from k-1 records. Our suppression and generalization based mechanisms can be viewed as instantiation of k-anonymity in the representation space. However, there are two major differences with traditional k-anonymity. First, our sensitive features represent non-sensitive columns from the point of view of k-anonymity literature and vice-versa. Second, sensitive columns are supposed to be categorical with high diversity in order to achieve good privacy, however, in our case sensitive columns (non-sensitive features) are unique to every individual and hence not offering identity protection explicitly. It is worth noting that under our threat model only sensitive attributes are supposed to be protected and our framework can provide protection against leakage of any arbitrary sensitive attribute including identity.