Adversarially Learned Representations for Information Obfuscation and Inference

Martin Bertran [*1] **Natalia Martinez** [*1] **Afroditi Papadaki** [2] **Qiang Qiu** [1] **Miguel Rodrigues** [2] **Galen Reeves** [1] **Guillermo Sapiro** [1]

Abstract

Data collection and sharing are pervasive aspects of modern society. This process can either be voluntary, as in the case of a person taking a facial image to unlock his/her phone, or incidental, such as traffic cameras collecting videos on pedestrians. An undesirable side effect of these processes is that shared data can carry information about attributes that users might consider as sensitive, even when such information is of limited use for the task. It is therefore desirable for both data collectors and users to design procedures that minimize sensitive information leakage. Balancing the competing objectives of providing meaningful individualized service levels and inference while obfuscating sensitive information is still an open problem. In this work, we take an information theoretic approach that is implemented as an unconstrained adversarial game between Deep Neural Networks in a principled, data-driven manner. This approach enables us to learn domain-preserving stochastic transformations that maintain performance on existing algorithms while minimizing sensitive information leakage.

1. Introduction

Information sharing and electronic communications permeate every aspect of human life. Shared data can contain information about many attributes, some of them are of interest for a particular task, while others can disclose irrelevant, conflicting, or sensitive information. As an example, a facial image contains information about features such as gender, emotion, ethnicity, and identity, among others. A user sharing an image of their face might be interested in a particular inference task (e.g., subject verification), but may want to remove some sensitive attribute (e.g., emotion recognition). As an additional use-case, imagine a subset of users wish to unlock their phone using facial identification, while others opt instead to verify their right to access the phone using other methods; in this setting, we wish the face identification service to collect information only on the consenting subset of users.

Here we address the problem of transactional information sharing, where a user discloses data about themselves in order to receive a service (e.g., identity-verification). Minimizing information leakage on a sensitive attribute (e.g., their emotion) while still providing a meaningful level of individualized service is still an open problem. There is significant prior work in related topics such as visual privacy through image obfuscation (McPherson et al., 2016; Oh et al., 2016; 2017; Brkic et al., 2017; Raval et al., 2017; Wu et al., 2018) and inpanting (Sun et al., 2018a;b; Orekondy et al., 2018), domain adaptation (Tzeng et al., 2017), protecting training data in machine learning (Shokri & Shmatikov, 2015; Zhang, 2018), fairness (Madras et al., 2018) and differential privacy (Dwork, 2008).

We introduce a learning framework based on mutual information to approach this challenge of balancing per-subject information obfuscation and utility preservation, where the data is sanitized prior to disclosure. There are theoretical works on studying utility-privacy tradeoffs using information metrics (Sankar et al., 2013; Basciftci et al., 2016), but, to our knowledge, none have been applied to high-dimensional data (e.g.: images) in a data-driven manner. The use of mutual information (MI) facilitates the theoretical analysis of performance bounds, which relates to important interpretable performance metrics such as accuracy (Feder & Merhav, 1994) and generalization error (Bassily et al., 2018). Unique aspects of the problem addressed here include: utility is measured per user, not as an aggregate statistic; no assumptions are made on the structure of the data, in particular, the utility and sensitive attribute might be strongly codependent and/or spatially co-localized. This differentiates us from related works on visual privacy (Sun et al., 2018a;b; Orekondy et al., 2018) where the utility is usually perceptual or semantical naturalness. Our sanitization (obfuscation) objective is set up as a constrained

[*]Equal contribution [1]Duke University, Durham, North Carolina, USA. [2]University College London, London, UK. Correspondence to: Martin Bertran <martin.bertran@duke.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

optimization problem, where different obfuscation and utility tradeoffs can be achieved, the sanitization transform is learned in a data-driven fashion using an adversarial approach with Deep Neural Networks (DNNs). Disadvantages of this approach include lack of naturalness in the resulting images, this is something that is not enforced explicitly but it could be incorporated as an additional utility objective. Another drawback is the reliance on auxiliary DNNs to measure Mutual Information.

1.1. Main Contributions and Manuscript Organization

We consider a scenario where a user wants to share a sanitized representation Y of high-dimensional data X in a way that a latent variable U can be inferred, but a sensitive latent variable S is obfuscated. X and Y can potentially be supported on the same domain, e.g., image-in-image-out.

In Section 2 we motivate the proposed framework as a distribution matching problem, and show that this can be formulated as a constrained optimization problem where both the objective function and the constraints are defined in terms of mutual information. Additionally, we derive bounds on the optimal performance of this proposed framework, and validate them on controlled and real experiments. We establish connections and comparisons between the proposed formulation and other important methodologies in related topics such as Information Bottleneck with Side Information (IBSI) (Chechik & Tishby, 2003; Chechik et al., 2005) and Differential Privacy (Dwork, 2008).

Section 3 shows how this optimization problem can be solved in a data-driven fashion by setting an adversarial game between competing DNNs. This formulation is used to learn *domain-preserving* data transformations that can accommodate for existing processing pipelines. That is, we can ensure that an existing algorithm that could be used to infer the utility variable U from the original data X can still be used to infer U from the filtered data Y that contains minimal (obfuscated) information on the sensitive attribute S.

Experiments on synthetic datasets are shown in Section 4. In Section 5, we exemplify the use of this framework on real data through the following use-cases: *Gender vs Emotion*, where emotion is obfuscated from the filtered image, but gender can still be inferred; *Subject vs Gender*, where face images are trained to retain subject verification performance while obfuscating gender inference; and *Subject vs Subject*, where the goal is to allow subject verification only on a subset of consenting users, non-consenting user's identities are obfuscated and made hard to recover from the filtered images. These examples measure both utility and sensitivity at the individual level, with both utility and sensitivity having varying degrees of dependence. Filtered images are domain-preserving, and can be used on an utility inference algorithm trained on the original data. Concluding remarks and future work are provided in Section 6.

2. Problem Formulation

We consider a scenario in which we have access to a set of three random variables X, U, and S, with joint distribution $p(x, u, s)$. Here $X \in \mathcal{X}$ is the data we observe (possibly high-dimensional), $U \in \mathcal{U}$ is a latent variable that we want to communicate (utility), and $S \in \mathcal{S}$ is a sensitive variable that we want to obfuscate. For the purposes of the analysis in this paper, we restrict ourselves to cases where $\mathcal{U}$ and $\mathcal{S}$ are finite alphabets. The advantage of this assumption, suitable for classification tasks, is that we ensure that the mutual informations of interest are always bounded. Unlike some other formulations (Chechik & Tishby, 2003; Chechik et al., 2005), we do not make any other assumption on $p(x, u, s)$, i.e., there is no underlying assumption that U, S and X define a Markov chain, or that U and S are conditionally independent given X, or that the distribution of X given U and S belong to any particular family.

Our goal is to find a stochastic transformation from X to a variable Y, $p(y \mid x)$, that provides information on U but not on S. Unlike the IBSI formulation, we do not explicitly minimize $I(X;Y)$. We want to find $p(y \mid x)$ such that the posterior distributions of the utility variable are similar given the filtered and original data, $p(u \mid y) \sim p(u \mid x)$, while the posterior of sensitivity variable S given obfuscated data Y is as close as possible to the prior, $p(s \mid y) \sim p(s)$ (meaning that observing Y does not change our beliefs about S). In many cases, both goals cannot be met simultaneously. However, we can formulate a problem where both objectives reach a compromise.

In the proposed formulation, we measure distances between distributions using KL divergence, $D_{KL}(p(u \mid x) \,\|\, p(u \mid y))$ and $D_{KL}(p(s \mid y) \,\|\, p(s))$ in particular. By taking the expectation of these metrics with respect to X and Y we recover the following mutual information:

$$\begin{aligned} E_{X,Y}[D_{KL}(p(u \mid x) \,\|\, p(u \mid y))] &= I(U; X \mid Y), \\ E_Y[D_{KL}(p(s \mid y) \,\|\, p(s))] &= I(S; Y). \end{aligned} \tag{1}$$

Both quantities have intuitive interpretations, $I(U; X \mid Y)$ is the amount of information on U we lose by observing the filtered data Y instead of the original data X, we call this quantity information loss (Geiger & Kubin, 2011). $I(S;Y)$ is the mutual information we disclose on variable S by observing variable Y, this is the information we fail to obfuscate. Under this setting, our objective is:

$$\min_{p(y|x)} \; I(U; X \mid Y) \quad s.t. \; I(S;Y) \leq k. \tag{2}$$

Here $k > 0$ is a constant controlling our tolerance on the amount of information on S disclosed via Y. Since U is conditionally independent on Y given X, $I(U; X \mid Y) = I(U;X) - I(U;Y)$, which leads to the the equivalent objective $\min_{p(y|x)} I(U; X \mid Y) \sim \max_{p(y|x)} I(U;Y)$. This

formulation is easier to connect with the Information Bottleneck (IB) approach (Tishby et al., 2000; Slonim & Tishby, 2000), which has been applied to other contexts in machine learning (see also (Achille & Soatto, 2017)). Figure 1 shows an illustrative Venn diagram of some of the quantities of interest in our formulation.

Note that in Eq.(1) we minimize $D_{KL}(p(u \mid x) \,\|\, p(u \mid y))$ and $D_{KL}(p(s \mid y) \,\|\, p(s))$ in expectation. This differs from other frameworks such as Differential Privacy (DP) (Dwork, 2008; Dwork et al., 2015; Rogers et al., 2016; Holohan et al., 2015) where guarantees are provided as a worst-case scenario, a strictly stronger notion than the one proposed here. Our formulation focuses on the problem of allowing inference on an attribute, while protecting inference of a different attribute as best as possible, this contrasts with the notion of protecting the anonymity of a data sample in its entirety. This makes our formulation suitable for a different set of tasks, particularly when we wish to infer attributes per data item (per user). Note that minimizing mutual information can lead to desirable characteristics such as tighter bounds on generalization error (Bassily et al., 2018; Xu & Raginsky, 2017; Asadi et al., 2018; Bousquet & Elisseeff, 2002; Nokleby et al., 2016).

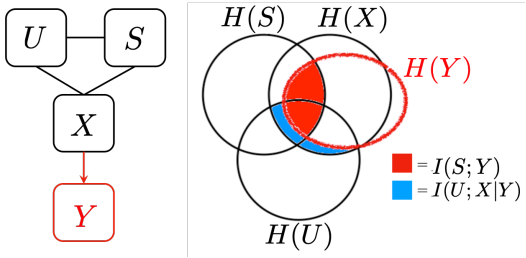

Figure 1. Left: Dependency graph of the observed variable X and the latent utility and sensitive variables U and S. Note that U and S can be codependent, thereby forcing a compromise between utility and obfuscation, this is addressed naturally via mutual information. Right: Venn diagram illustrating conditional mutual information that constrains the performance of any sanitization mapping $Y \sim p(y \mid x)$ such that $(Y \perp\!\!\!\perp (U,S)) \mid X$. The information leakage $I(S;Y)$ and censured information $I(U;X \mid Y)$ shown in red and blue respectively cannot always be simultaneously set to 0, since they are partially at odds.

To solve this constrained optimization problem, we relax it using the quadratic penalty method (Wright & Nocedal, 1999), which penalizes solutions that violate the desired constraints. Since $I(U;X \mid Y)$ and $I(S;Y)$ are both continuously differentiable functions of $p(y \mid x)$, this relaxed problem is equivalent to the constrained one as $\lambda \to \infty$:

$$\min_{p(y|x)} \left[I(U;X \mid Y) + \lambda \max(I(S;Y) - k, 0)^2 \right]. \tag{3}$$

A counterexample on why a more intuitive loss function $I(U;X \mid Y) + \lambda I(S;Y)$ is not used is shown in Supplementary Material.

Section 3 details how to implement Eq.(3) as an adversarial game between competing DNNs using standard training tools. Before that, Section 2.1 shows that the best possible performance for the global optimal solution to Eq.(2) can be lower-bounded with terms that depend only on mutual informations and the joint distribution $p(u,s)$. For categorical variables (e.g., classification tasks), this joint distribution can be easily computed from the observed contingency tables.

2.1. Performance Bounds

One key question that arises is what are the intrinsic limits on the trade-offs attainable from the solutions to Eq.(2) for any given problem. In this section we provide lower and upper bounds that give insight into feasible solutions. In particular, a large gap between upper and lower bounds suggest the existence of solutions with good utility-obfuscation trade-offs. Bounds presented in this section are also valid for continuous X, U, S as long as $I(U;X), I(S;Y)$ are bounded, but for clarity we derive them for discrete variables X, U, S.

In many real-world classification problems, the support of the observed variable X is high-dimensional, while the support over the latent variables of interest U and S is comparatively smaller. Lemma 2.1, presented next, shows that we can bound the solution of Eq.(2) by considering mappings that go directly from the latent variables U and S to the obfuscated variable Y. This simplifies the analysis since $|\mathcal{U} \times \mathcal{S}| \ll |\mathcal{X}|$ for many problems of interest.

Lemma 2.1. Let U and S be a pair of latent variables, drawn from $p(u,s)$ and supported on finite alphabets $\mathcal{U}$ and $\mathcal{S}$ respectively, and let X be the observed variable drawn from the conditional distribution $p(x \mid u,s)$ supported on $\mathcal{X}$. Let $Y \in \mathcal{Y}$ be jointly distributed with U, S, and X. The following relation holds:

$$\begin{aligned} \min_{p(y|x)\in\Omega} I(U;X \mid Y) \geq \min_{p(y|u,s)\in\Omega^*} I(U;X) - I(U;Y), \\ \Omega = \{p(y|x) : p(y|x) = p(y|x,u,s); \quad I(S;Y) \leq k\}, \\ \Omega^* = \{p(y|u,s) : I(U;Y) \leq I(U;X); \quad I(S;Y) \leq k\}. \end{aligned} \tag{4}$$

Proof: In the left term of the inequality $(U,S) \to X \to Y$ form a markov chain, by data processing inequality $\Omega = \Omega \cap \{p(y|x) : I(U;Y) \leq I(U;X)\}$ then:

$$\begin{aligned} \min_{p(y|x)\in\Omega} I(U;X \mid Y) &= \min_{p(y|x)\in\Omega} I(U;X) - I(U;Y), \\ &\geq \min_{p(y|u,s)\in\Omega^*} I(U;X) - I(U;Y) \,\square\,. \end{aligned} \tag{5}$$

The main idea in Lemma 2.1 is that we can obtain lower bounds without worrying about the details of $p(x|u,s)$. However, evaluating Eq.(4) can still be a challenging problem on itself, particularly when $|\mathcal{Y}|$ is large. To analyze this, we can obtain a sequence of upper bounds to Eq.(4) where we constrain the cardinality of $|\mathcal{Y}|$ to be finite. We will refer

to this as the restricted cardinality sequence (RCS). Each element of this sequence requires optimization over a finite transition matrix $p(y \mid u, s)$. Computation details are shown in Supplementary Material.

An alternative lower bound that can prove to be easier to compute is provided in Lemma 2.2. This lower bound only requires the computation of $I(U;S)$ and $I(U;S \mid X)$.

Lemma 2.2. Let X, U, S be three discrete random variables with joint probability distribution $p(x,u,s)$. For any solution to Eq.(2) with tolerance k we have

$$\begin{aligned} I(U;X \mid Y) \; &\geq -I(S;Y) + I(U;S) - I(U;S \mid X), \\ &\geq -k + I(U;S) - I(U;S \mid X). \end{aligned} \tag{6}$$

Proof: Consider the following equality, derived from properties of Shannon Information, see (Yeung, 2012).

$$\begin{aligned} I(U;S) - I(U;S \mid X) \; &= I(S;Y) - I(S;Y \mid U) \\ &+ I(U;X \mid Y) - I(U;X \mid Y,S). \end{aligned} \tag{7}$$

A complete proof of this equality is provided in Supplementary Material for completeness. The stated inequality follows from $I(S;Y) \leq k$ and the non-negativity of mutual information $\square$.

Finally, Lemma 2.3 provides an achievable upper bound to the solution to Eq.(2).

Lemma 2.3. Let X, U, S be three discrete random variables with joint probability distribution $p(x,u,s)$. For all $k > 0$ There exists a conditional distribution $p(y \mid x)$ such that

$$\begin{aligned} &I(S;Y) \leq k, \\ &I(U;X \mid Y) = \max(0, 1 - \tfrac{k}{I(S;X)}) I(U;X) \end{aligned} \tag{8}$$

Proof: $\forall k > 0$, let $\beta = \min(\frac{k}{I(S;X)}, 1) \in [0,1]$. Consider $p(y \mid x) = \beta\delta(y = x) + (1-\beta)\delta(y = \xi)$, where $\xi \notin \mathcal{X}$. Note that $H(S \mid Y = \xi) = H(S)$ and $H(S \mid Y \neq \xi) = H(S \mid X)$. Therefore

$$\begin{aligned} I(S;Y) &= H(S) - H(S \mid Y), \\ &= H(S) - \beta H(S \mid Y \neq \xi) - (1-\beta) H(S \mid Y = \xi), \\ &= \beta(H(S) - H(S \mid X)) = \beta I(S;X). \end{aligned} \tag{9}$$

Analogously $I(U;Y) = \beta I(U;X)$. The statement of the lemma follows from $I(U;X \mid Y) = I(U;X) - I(U;Y)$.$\square$

2.2. Analysis of Lower Bounds

Lemmas 2.2 and 2.1 provided two lower bounds to Eq.(2). Note that Lemma 2.2 provides a linear lower bound, while Lemma 2.1 has no such restriction. Here we show two examples where the RCS approximation to Lemma 2.1 seems to converge in a small number of iterations, and the bound it suggests is comparatively tighter than the one provided by Lemma 2.2.

Figure 2 shows the lower bound derived in Lemma 2.2, and elements of the restricted cardinality sequence (RCS) used to approximate Lemma 2.1 for a particular choice of distribution of latent and observed variable (U, S, X). Note that the infimum of the RCS sequence is a true lower bound to Eq.(2).

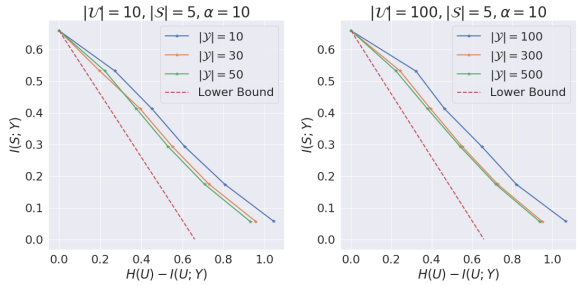

Figure 2. Lower bound derived in Lemma 2.2 and elements of the RCS sequence are shown for two particular choices of joint distributions. Left figure shows trade-offs when $|\mathcal{U}| = 10, |\mathcal{S}| = 5$, while right figure shows comparable trade-offs for $|\mathcal{U}| = 100, |\mathcal{S}| = 5$. The conditional distribution generating both datasets is $P(u,s) \propto 1 + \alpha\delta(u \bmod |\mathcal{S}| = s)$, where $\alpha > 0$ is chosen to control $I(U;S)$. Images were produced with $\alpha = 10$. Figure also shows that the RCS sequence quickly converges to a limiting lower bound.

3. Data-Driven Implementation

Even if the joint distribution of $P_{X,U,S}$ is not known, it is possible to implement Eq.(3) in a data-driven fashion to find $p(y \mid x)$ in a family of parametric stochastic neural network architectures $q_\theta(x,z) : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$. We illustrate this now.

Let $Z \sim P_Z$ be a random variable drawn from a known distribution, and let θ be the network parameters of the transformation we wish to learn. Note that $y = q_\theta(x,z)$ is a deterministic value for any pair (x,z), but $Y_\theta = q_\theta(x,Z)$ is a random variable drawn from an implicit conditional distribution $p_\theta(y \mid x)$.

Assume we have access to a labeled dataset $\{(x_i, s_i, u_i)\}_{i=1}^N$, where s_i and u_i are the true values of the secret and utility variables for observation x_i. Learning a parametric stochastic representation $Y = q_\theta(x,z)$ that optimizes Eq.(3) requires estimating the posteriors: $P_{S|Y}$, $P_{U|Y}$, and $P_{U|X}$; these estimators are obtained through parametric neural networks $p_\eta(s \mid y)$, $p_\psi(u \mid y)$, and $p_\phi(u \mid x)$ respectively. Under this setup q_θ is obtained by simultaneously optimizing the following adversarial objectives:

$$\begin{aligned} &\hat{\eta} = \text{argmin}_\eta E_{X,S,Z}\big[-log(p_\eta(s \mid q_{\hat{\theta}}(x,z))\big], \\ &\hat{\psi} = \text{argmin}_\psi E_{X,U,Z}\big[-log(p_\psi(u \mid q_{\hat{\theta}}(x,z))\big], \\ &\hat{\phi} = \text{argmin}_\phi E_{X,U}\big[-log(p_\phi(u \mid x)\big], \\ &\hat{\theta} = \text{argmin}_\theta E_{X,Z}\big[D_{KL}(p_{\hat{\phi}}(u \mid x) \,\|\, p_{\hat{\psi}}(u \mid q_\theta(x,z)))\big] \\ &\quad + \lambda \max(E_{X,Z}\big[D_{KL}(P_s \,\|\, p_{\hat{\eta}}(s \mid q_\theta(x,z)))\big] - k, 0)^2. \end{aligned} \tag{10}$$

The first three equations in Eq.(10) are cross-entropy loss terms to ensure the estimators $p_\eta(s \mid q_\theta)$, $p_\psi(u \mid q_\theta)$, and $p_\phi(u \mid x)$ are all good estimators to the true posterior distributions. The last loss term is a direct translation of Eq.(3). Details on the algorithmic implementation are shown in Algorithm 1.

Algorithm 1 Adversarial Information Obfuscation

Input: data $\{(x_i, s_i, u_i)\}$; hyperparameters (lr, λ, k)
$P(s)$ is the empirical marginal distribution of $\{s_i\}$
repeat
  Draw b samples from dataset
  $(x_{(1)}, u_{(1)}, s_{(1)}), ...(x_{(b)}, u_{(b)}, s_{(b)}) \sim P_{X,U,S}$
  Draw b samples from sampling distribution
  $z_{(1)}, ...z_{(b)} \sim P_Z$
  Evaluate cross-entropy loss on posterior inference networks:
  $\Phi(\phi) = \frac{1}{b}\sum_{i=1}^{b} -logP_\phi(u_{(i)} \mid x_{(i)})$
  $\Psi(\psi) = \frac{1}{b}\sum_{i=1}^{b} -logP_\psi(u_{(i)} \mid Q_\theta(x_{(i)}, z_{(i)}))$
  $H(\eta) = \frac{1}{b}\sum_{i=1}^{b} -logP_\eta(s_{(i)} \mid Q_\theta(x_{(i)}, z_{(i)}))$
  Stochastic gradient descent:
  $\phi \leftarrow \phi - lr\nabla_\phi\Phi(\phi)$; $\psi \leftarrow \psi - lr\nabla_\psi\Psi(\psi)$
  $\eta \leftarrow \eta - lr\nabla_\eta H(\eta)$
  Evaluate unconstrained penalty loss:
  $\Theta(\theta) = \frac{1}{b}\sum_{i=1}^{b} log\frac{P_\phi(u_{(i)}|x_{(i)})}{P_\psi(u_{(i)}|Q_\theta(x_{(i)},z_{(i)}))} + \lambda\max(\frac{1}{b}\sum_{i=1}^{b} log\frac{P_\eta(s_{(i)}|Q_\theta(x_{(i)},z_{(i)}))}{P(s_{(i)})} - k, 0)^2$
  Stochastic gradient descent:
  $\theta \leftarrow \theta - lr\nabla_\theta\Theta(\theta)$
until Convergence

3.1. Domain-Preserving Transformations with Fixed Utility Inference Algorithms

An attractive application of the proposed method arises when we impose two additional restrictions. We first constrain transformations to be strictly domain preserving (e.g., an image to image transformation). The second constraint is that we are given a fixed utility inference algorithm $p_{\hat{\phi}}(u \mid \cdot)$ that works on unfiltered, original data X and cannot be modified. Therefore, we require that this algorithm also performs well on the filtered data Y. These additional requirements allow us to integrate sensitivity constraints on an existing data-processing pipeline with minimal or no disruption.

These objectives can be accomplished with a small variant of the adversarial training described in Algorithm 1, where $p_{\hat{\phi}}(u \mid \cdot)$ is used in place of $p_{\hat{\psi}}(u \mid \cdot)$ to evaluate the unconstrained penalty objective $\Theta(\theta)$, and where no stochastic gradient descent update step is performed on the parameters ϕ. This Algorithm is further detailed in Supplementary Material. Figure 3 shows a schematic representation of this particular use-case.

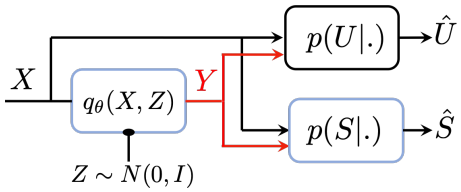

Figure 3. Schematic of the three main components in the adversarial information obfuscation framework with a fixed utility algorithm. Original data X can be directly fed into the algorithms that infer the sensitive information ($p(S \mid \cdot)$) and the utility ($p(U \mid \cdot)$). Since the mapping $Y = q_\theta(X, Z)$ is domain preserving ($Y \in \mathcal{X}$), the filtered data can also be directly fed to both tasks without any need for further adaptations.

4. Validation on Synthetic Data

We first study the performance of the proposed framework on synthetic data and compare the results with the performance bounds derived in Section 2.1. Here we show that the proposed formulation is able to obtain nontrivial trade-offs for both a standard neural network and a linear filter.

For these experiments, the observed data is a perfect representation of the latent variables, $X = (U, S)$, U is uniformly distributed on $|\mathcal{U}| = 6$, and S is uniformly distributed on $|\mathcal{S}| = 2$. The design parameter for the model is the mutual information $I(U; S)$, which can take values in the range $[0, ln(2)]$. We compare the tradeoff curves for $I(U; X \mid Y)$ and $I(S; Y)$ for two simple classes of stochastic obfuscation transforms, a linear filter $Y = AX + Z, Z \sim N(0, \sigma^2)$; and a stochastic neural network with less than 300 parameters total. The posterior inference networks $p_\eta(S \mid \cdot)$, $p_\psi(U \mid \cdot)$, and $p_\phi(U \mid \cdot)$ are implemented using fully connected neural networks. Detailed descriptions of the architectures, data generating process, and design parameters are provided in Supplementary Material.

Figure 4 summarizes the trade-offs obtained by both filter classes for three levels of codependence $I(U; S) = [0.2, 0.4, 0.69] nat$ (nat is the natural unit of information), The bounds presented in Section 2.1 are also shown for reference. The nonlinear filter performs slightly better than the linear filter. The point $(I(U; X \mid Y) = 0, I(S; Y) = I(U; S))$ is always reachable under this generation model, and corresponds to communicating U perfectly through Y, while blocking any other "direct" observation of S; both linear and nonlinear filters easily achieve this tradeoff point.

We note that the results obtained by both filters, which rely on data-driven optimization, are reasonable approximations to the bounding RCS sequence found by directly optimizing over transition probabilities described in Lemma 2.1. The performance gap can be understood as a failure of the filters to explore different data representation modalities, something we believe to be a related to the problem of mode collapse, which is commonly observed and studied in

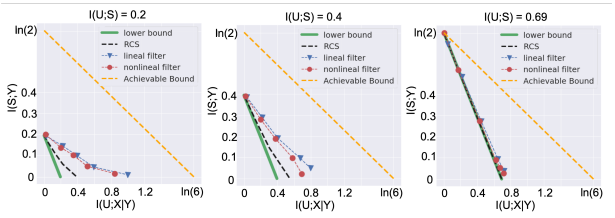

Figure 4. Tradeoff curves in the $I(U;X \mid Y), I(S;Y)$ plane for linear and nonlinear filters, columns represent different levels of mutual information in the joint distribution of (U,S), $I(U;S) = [0.2, 0.4, 0.69]nat$. The theoretical bounds derived in Section 2.1 are shown for comparison. High tolerance (high $I(S;Y)$) points are easier to achieve.

GANs, see for example (Metz et al., 2016; Salimans et al., 2016; Wu et al., 2018; Arjovsky et al., 2017; Srivastava et al., 2017).In supplementary material we show some of the learned representations Y on the linear and nonlinear architectures, and the RCS sequence from Lemma 2.1.

5. Results on Facial Images

The following example are based on the framework presented in Section 3.1, shown in Figure 3. Since the utility estimation algorithm ($p_{\hat{\phi}}(u \mid \cdot)$) is fixed, the learned mapping q_θ needs to be *domain-preserving* and it additionally needs to perform well when filtered data is fed through $p_{\hat{\phi}}(u \mid \cdot)$. This is an additional constraint that was not specifically modeled in Eq.(3), and is what enables a utility provider to be compatible with this sensitivity-obfuscating transformation without modifying existing algorithms. Note, however that the sensitive attribute estimator $p_\eta(s \mid \cdot)$ is always trained adversarially against the learned mapping q_θ.

First we tackle the problem of learning a mapping that preserves gender detection while hiding emotion (*emotion-and-gender*). We then learn a filter over facial images that makes subject verification possible only for a consenting subset of users (*subject-within-subject*). Finally, we show how we can disable gender recognition while allowing subject verification (*subject-and-gender*) a challenging task since the utility variable U is harder to infer than S, and they are both tightly coupled. All problems were implemented using the same network architectures, for the inference networks $p_\phi(u \mid \cdot)$ and $p_\eta(s \mid \cdot)$ we used Xception networks (Chollet, 2017), and for the learned filter $q_\theta(x, z)$ we used a U-Net-based network (Ronneberger et al., 2015). The stochastic component was added as an extra noise input appended to the network. Detailed architectures for both networks are shown in Supplementary Material.

5.1. Obfuscating Emotion While Preserving Gender

We apply the proposed adversarial information obfuscation framework to facial images X, the utility variable U is the gender of the imaged subject, while the sensitive attribute S is their emotion (smiling/non-smiling). In this scenario, variables U and S are almost independent. We conduct this experiment over the CelebA dataset (Liu et al., 2015). Table 1 shows the distribution of the utility and secrecy estimators over the sanitized data. A comparison against blacking out the face and blacking out the lower half of the face are also shown. Figure 5 shows examples of sanitized images. It is visually possible to identify the gender of the subject but not their emotion. Most importantly, the existing gender detection algorithm still performs correctly over the sanitized images.

Figure 5. Learned transformation to obfuscate emotion recognition and preserve gender recognition. Each column shows the same individual. Top row shows the original image, middle and bottom row show the filtered images for sensitive information budgets $0.5nat$ and $0.4nat$ respectively. The learned filtering behaviour is easy to interpret, the mouth area is increasingly blurred out. This behaviour was learned via the adversarial training regime described in Algorithm 1, and was not designed by the user.

Table 1. Results across several tolerance parameters k. The confidence of an adversarially trained emotion estimator is shown alongside an estimator trained only on natural images. Accuracy is also shown for the fully fixed gender detection network. Results on images where the lower face and full face are blacked out (LFB, FFB) are shown as comparison. Guessing values based on prior distribution are also shown

TOL k	FIXED EMOTION		ADV EMOTION		FIXED GENDER	
	CONF	ACC	CONF	ACC	CONF	ACC
∞	0.88	91.8%	0.88	91.8%	0.92	94.9%
0.5	0.65	68.4%	0.87	91.4%	0.86	89.3%
0.4	0.57	58.6%	0.81	85.8%	0.84	88.0%
0.3	0.55	56.8%	0.75	81.5%	0.83	86.7%
0.2	0.51	51.9%	0.67	74.3%	0.79	83.9%
LFB	0.55	53.9%	0.65	74.0%	0.72	75.5%
FFB	0.52	52.2%	0.62	71.5%	0.63	67.9%
GUESS	0.50	51.9%	-	-	0.52	60.7%

The results show that the adversarial information obfuscation algorithm learned a natural action to block emotion recognition, effectively driving the confidence of the emotion-inferring algorithm towards random guessing, this is observed on both the original emotion inference algorithm, as well as the adversarially trained one.

5.2. Subject within Subject

We now analyze the *subject-within-subject* problem. Here, only a small subset of potential users wish to use facial subject verification, the remaining users opt out of the feature and wish that their identity is obfuscated from the collected images. In this setting, the utility and sensitive variables can be thought of as belonging to mutually disjoint subsets of a latent variable.

We solve this problem by training a domain-preserving stochastic mapping q_θ on facial image data X, where the utility and secret variable U and S are categorical variables over consenting and non-consenting users respectively. We test this over the FaceScrub dataset (Kemelmacher-Shlizerman et al., 2016), again using Xception networks (Chollet, 2017) as the utility and secrecy inferring algorithm. The stochastic mapping was implemented using a stochastic adaptation of U-Net (Ronneberger et al., 2015), architecture details are provided in Supplementary Material.

Table 2 shows the top-5 categorical accuracy of the utility network over the sanitized data at various k points in the sensitivity-utility trade-off. Figure 6 show some representative images on how images are sanitized. It also shows that the sanitization function is able to preserve information about the utility variable while effectively censoring the secret variable, even for unobserved images.

Table 2. Top-5 accuracy performance of the subject detector after obfuscating the identity of non-consenting users for various tolerance levels k. Performance is shown across 3 subsets, consenting users are users that decided to be detected by the utility algorithm, observed private users are those that explicitly decided to protect their privacy, while unobserved private users are users that decided to protect their privacy but were not available during training. Consenting users are still recognized by the system, while non-consenting users are not.

TOLERANCE k	CONSENTING USER	OBS. PRIVATE USER	UNOBS. PRIVATE USER
∞	98.7%	98.4%	97.9%
3	98.3%	7.81%	9.38%
1	97.8%	4.69%	6.25%
0.5	97.6%	3.12%	4.69%
GUESS	2.5%	2.5%	2.5%

5.3. Preserving Subject Verification, Obfuscating Gender

Finally, we tackle the issue of preserving subject verification while obfuscating the gender, we do this on 200 subjects from the FaceScrub dataset (Kemelmacher-Shlizerman et al., 2016). This is a hard task, since the amount of information required to identify the gender is substantially smaller than the one necessary to perform subject verification. Furthermore, these two variables are strongly codependent ($I(U;S) \simeq \min(H(U), H(S))$). Illustrative examples of filtered images are shown on Figure 7

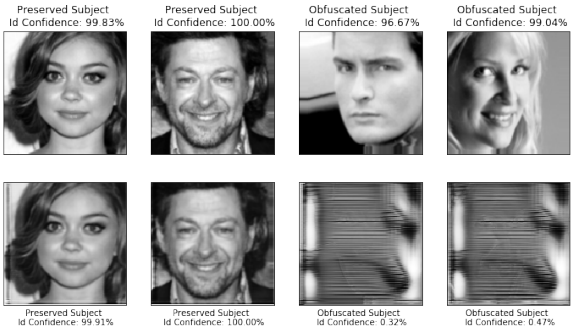

Figure 6. Left and right figures show images of consenting and non-consenting (private) users respectively, along with their sanitized counterparts. The identity of consenting users is still easily verified, while the identity of non-consenting users is effectively censored.

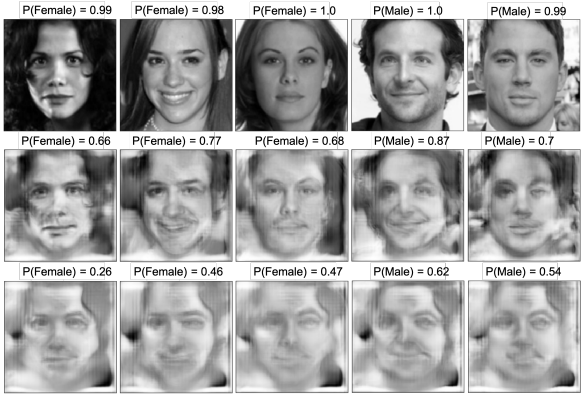

Figure 7. First row shows original images, second and third show the filtered images for sensitive information tolerances ($I(S;Y) < k$) 0.5*nat* and 0.3*nat* respectively. Gender probabilities where computed over 5 realizations of the filter. It is hard to differentiate genders as the tolerance decreases. Subject verification is still performed correctly, without altering the original algorithm, on all images shown. The learned filter was trained against an *adversarial* gender inference network, while the subject verification algorithm remained fixed.

The *subject-vs-gender* task is made harder by the fact that the subject identification algorithm was trained on natural images and remains fixed during training, while the gender inference algorithm is trained adversarially. For this reason, Table 3 additionally shows the Top-5 categorical accuracy obtained when the subject identification network is also allowed to retrain its final dense layer. Note that this still implies minimal changes to the subject inference network, since all lower feature extraction layers remain unchanged. As a baseline comparison, we show the performance of the Fader Network (Lample et al., 2017) gender mixing model; here images have their gender attribute rewritten at random. We note that for similar Top-5 accuracy in subject detection, our method significantly lowers gender detection accuracy

(~ 60% vs 88% on fixed gender detection, ~ 90% vs 96% on adversarial gender detection).

Table 3. Results across several tolerance parameters k. The confidence of an adversarially trained gender estimator is shown alongside an estimator trained only on natural images. Top-5 accuracy results are shown for both the fully fixed subject verification network, as well as one where only the final dense layer was retrained. The performance of randomly resampling the gender attribute using Fader Network (FaderN) and that of guessing the attribute based on prior information (Guess) are also shown for comparison.

TOL k	FIXED GENDER		ADV GENDER		FIXED SUBJECT	TRAINED SUBJECT
	CONF	ACC	CONF	ACC	TOP-5 ACC	TOP-5 ACC
∞	0.98	98.6%	0.98	98.6%	98.8%	98.8%
0.5	0.59	59.5%	0.86	90.2%	93.5%	96.8%
0.4	0.59	60.3%	0.80	85.3%	88.1%	94.9%
0.3	0.54	54.0%	0.72	79.4%	81.4%	92.8%
0.2	0.55	56.1%	0.67	74.6%	81.6%	91.0%
0.1	0.52	51.6%	0.59	67.1%	74.5%	89.6%
FADERN	0.87	87.8%	0.95	95.9%	92.5%	95.2%
GUESS	0.50	54.8%	-	-	2.5%	-

5.4. Performance bounds on Real Experiments

To conclude, Figure 8 compares the trade-offs achieved on both the *subject-vs-gender* and *gender-vs-emotion* (empricial tradeofs are computed by averaging across test-set images) against the lower bounds derived in Lemma 2.1 computed using the RCS approximations described in Supplementary Material (the joint distribution $p(u, s)$ required by the RCS is estimated from the label contingency table), and the upper bound derived in Lemma 2.3.

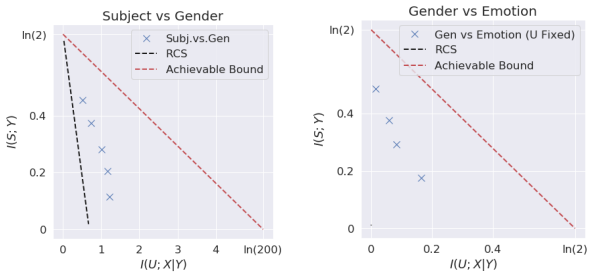

Figure 8. Trade-offs obtained via Algorithm 1 in the *subject-vs-gender* (left) and *gender-vs-emotion* (right). The results obtained are compared with the bound derived in Lemma 2.1 and approximated using the RCS method described in Supplementary Material. The achievable bound derived in Lemma 2.3 is also shown for reference.

6. Concluding Remarks

We addressed the problem of learning data representations that simultaneously obfuscate information about sensitive latent attributes, while preserving information about attributes we specifically wish to disclose (utility). This was formulated as a distribution matching problem, and we used tools from information theory to formalize this notion into a concrete optimization problem.

We derived easy-to-compute bounds on the optimal achievable performance of these transformations, and showed how the original constrained optimization problem has an equivalent unconstrained formulation that can be directly optimized as an adversarial game played between DNNs. We expanded the restrictions imposed on the problem by limiting ourselves to domain-preserving transformations (e.g., images to images), that preserve utility inference capabilities on an existing system, while defending against an adversarial network attempting to infer the sensitive attribute from the learned representation.

Experimental results show that the learned representations perform well when compared against the theoretically-derived bounds; the performance differential can be potentially interpreted as a behaviour akin to mode collapse in GANs (Goodfellow et al., 2014).

Results on facial image data show that the framework is able to handle hard-to-model tasks such as hiding emotion recognition while enabling gender identification; and preserving subject verification capabilities on a subset of individuals, while disallowing this on non-consenting individuals. We also showed excellent performance on the challenging task of preserving identity while obfuscating gender. The learned behaviour of the filters trained on these facial images varied significantly from task to task, but was altogether interpretable.

It is important to highlight that the filters learned through this adversarial framework are as good as the estimators for the utility and secret variables allow them to be, this is especially true for the mutual information estimators. A future challenge to address is how to get consistent estimators that perform well under the type of shifting inputs that the filter produces, one of the multiple novel challenges presented by these types of approaches.

An implementation of this framework is available at `www.github.com/MartinBertran/AIOI`.

Acknowledgements

Work partially supported bu DoD, NSF, NIH, Cisco, Microsoft, and Amazon.

References

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50), 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Asadi, A. R., Abbe, E., and Verdú, S. Chaining mutual

information and tightening generalization bounds. *arXiv preprint arXiv:1806.03803*, 2018.

Basciftci, Y. O., Wang, Y., and Ishwar, P. On privacy-utility tradeoffs for constrained data release mechanisms. In *2016 Information Theory and Applications Workshop (ITA)*, pp. 1–6. IEEE, 2016.

Bassily, R., Moran, S., Nachum, I., Shafer, J., and Yehudayoff, A. Learners that use little information. In *Algorithmic Learning Theory*, pp. 25–55, 2018.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Brkic, K., Sikiric, I., Hrkac, T., and Kalafatic, Z. I know that person: Generative full body and face de-identification of people in images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1319–1328. IEEE, 2017.

Chechik, G. and Tishby, N. Extracting relevant structures with side information. In *Advances in Neural Information Processing Systems*, pp. 881–888, 2003.

Chechik, G., Globerson, A., Tishby, N., and Weiss, Y. Information bottleneck for gaussian variables. *Journal of machine learning research*, 6(Jan):165–188, 2005.

Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 1800–1807. IEEE, 2017.

Dwork, C. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pp. 1–19. Springer, 2008.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pp. 2350–2358, 2015.

Feder, M. and Merhav, N. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1):259–266, 1994.

Geiger, B. C. and Kubin, G. On the information loss in memoryless systems: The multivariate case. *arXiv preprint arXiv:1109.4856*, 2011.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Holohan, N., Leith, D. J., and Mason, O. Differential privacy in metric spaces: Numerical, categorical and functional data under the one roof. *Information Sciences*, 305:256–268, 2015.

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., and Brossard, E. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4873–4882, 2016.

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pp. 5967–5976, 2017.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

McPherson, R., Shokri, R., and Shmatikov, V. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Nokleby, M., Beirami, A., and Calderbank, R. Rate-distortion bounds on bayes risk in supervised learning. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pp. 2099–2103. IEEE, 2016.

Oh, S. J., Benenson, R., Fritz, M., and Schiele, B. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2016.

Oh, S. J., Fritz, M., and Schiele, B. Adversarial image perturbation for privacy protection a game theory perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1491–1500. IEEE, 2017.

Orekondy, T., Fritz, M., and Schiele, B. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8466–8475, 2018.

Raval, N., Machanavajjhala, A., and Cox, L. P. Protecting visual secrets using adversarial nets. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1329–1332. IEEE, 2017.

Rogers, R., Roth, A., Smith, A., and Thakkar, O. Max-information, differential privacy, and post-selection hypothesis testing. *arXiv preprint arXiv:1604.03924*, 2016.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and*

computer-assisted intervention, pp. 234–241. Springer, 2015.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Sankar, L., Rajagopalan, S. R., and Poor, H. V. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852, 2013.

Shokri, R. and Shmatikov, V. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321. ACM, 2015.

Slonim, N. and Tishby, N. Agglomerative information bottleneck. In *Advances in neural information processing systems*, pp. 617–623, 2000.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., and Sutton, C. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pp. 3308–3318, 2017.

Sun, Q., Ma, L., Joon Oh, S., Van Gool, L., Schiele, B., and Fritz, M. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5050–5059, 2018a.

Sun, Q., Tewari, A., Xu, W., Fritz, M., Theobalt, C., and Schiele, B. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 553–569, 2018b.

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

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

Wright, S. and Nocedal, J. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

Wu, Z., Wang, Z., Wang, Z., and Jin, H. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 606–624, 2018.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 2524–2533, 2017.

Yeung, R. W. *A first course in information theory*. Springer Science & Business Media, 2012.

Zhang, T. Privacy-preserving machine learning through data obfuscation. *arXiv preprint arXiv:1807.01860*, 2018.