Obfuscation via Information Density Estimation

Hsiang Hsu
Harvard University
hsianghsu@g.harvard.edu

Shahab Asoodeh
Harvard University
shahab@seas.harvard.edu

Flavio P. Calmon
Harvard University
flavio@seas.harvard.edu

Abstract

Identifying features that leak information about sensitive attributes is a key challenge in the design of information obfuscation mechanisms. In this paper, we propose a framework to identify information-leaking features via information density estimation. Here, features whose information densities exceed a pre-defined threshold are deemed information-leaking features. Once these features are identified, we sequentially pass them through a targeted obfuscation mechanism with a provable leakage guarantee in terms of E_γ-divergence. The core of this mechanism relies on a data-driven estimate of the trimmed information density for which we propose a novel estimator, named the *trimmed information density estimator* (TIDE). We then use TIDE to implement our mechanism on three real-world datasets. Our approach can be used as a data-driven pipeline for designing obfuscation mechanisms targeting specific features.

1 Introduction

A challenging problem in dataset and information sharing platforms is limiting the leakage of sensitive or private information. Sensitive information leakage can be controlled by *obfuscating* samples in a dataset prior to disclosure; i.e., perturbing the sample in a way that sensitive information cannot be effectively inferred (Bertran et al., 2019; Chen et al., 2019; Zemel et al., 2013). Samples may contain several *features*, only some of which might leak information about sensitive attributes. For example, not all areas in a facial image equally disclose emotion (as a sensitive attribute), and not all terms used in Tweets equally reveal a user's political preference. Given a set of sensitive attributes, an information obfuscation mechanism should ideally target only those features of the data that leak excessive amount of sensitive information. Such mechanisms usually achieve higher utility (e.g., the quality of the image) by incorporating either complete (cf. information-theoretic privacy (Calmon and Fawaz, 2012; Asoodeh et al., 2018; Issa et al., 2018; Hsu et al., 2018; Diaz et al., 2018)) or partial (cf. generative adversarial privacy (Huang et al., 2017)) knowledge of the underlying data distribution.

In this paper, we propose a data-driven information-obfuscation mechanism. As a natural first step, we *identify* the information-leaking features in the data via an information-theoretic quantity called the *information density* (Pinsker, 1964). This quantity is at the heart of most information-theoretic measures of privacy (Asoodeh et al., 2018; Issa et al., 2018; Hsu et al., 2018) as well as differential privacy (DP) (Bun and Steinke, 2016; Dwork and Rothblum, 2016; Balle and Wang, 2018; Chaudhuri et al., 2011). Intuitively, the information density captures the change of the belief about a sensitive attribute upon an observation of a sample in a disclosed dataset.

Features whose information density are above a certain threshold (which we call information-leaking features) can be randomized (e.g., perturbed) via an obfuscation mechanism. The goal of the obfuscation mechanism is to limit unwanted inferences about a sensitive attribute from disclosed data. We argue that this objective can be mathematically formulated in terms of a specific type of f-divergence (Csiszár, 1967), called the E_γ-divergence, which captures the tail distribution of the information density. We propose a feature-dependent Gaussian mechanism that ensures obfuscation in terms of E_γ-divergence by targeting only the information-leaking features.

The methodology proposed here aims to develop a theoretical foundation for expounding existing approaches that completely rely on neural networks to identify and obfuscate the information-leaking features (Bertran et al., 2019; Chen et al., 2019). Despite its theoretical nature, our approach has a comparable performance in terms of sensitive information leakage as Bertran et al. (2019), without a specific "utility" target having to be pre-determined by a user. Furthermore, it adds a layer of interpretability, enabling features that pose an ex-

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

cessive leakage risk to be identified and communicated to the data owner.

In practice, we need to estimate the information density from samples. This estimation problem is inherently connected to mutual information estimation (since the expected value of information density is equal to the mutual information) which is known to be challenging (Valiant and Valiant, 2011; Wu and Yang, 2016; Gao et al., 2017) unless an adequate parametric model is assumed (Vapnik, 2013). The main difficulty lies in the unboundedness of the information density, which leads to high sample complexity for reliable estimation. However, since our mechanism perturbs only information-leaking features, it requires the *trimmed information density* whose estimation is a much easier task than the original information density estimation problem. Inspired by Belghazi et al. (2018); Liu et al. (2017), we develop the trimmed information density estimator (TIDE), based on the variational representations of f-divergences (Nguyen et al., 2010).

The contributions of this paper, from theoretical results to practice, are listed as follows:

1. We propose a framework for identifying information-leaking features by the trimmed information density, and use the E_γ-divergence between the distributions over a sensitive attribute prior and posterior to a disclosed sample to measure the information leakage. Moreover, we show that obfuscation mechanisms that aim to minimize the E_γ-divergence satisfies several desirable properties in terms of information leakage guarantees (cf. Section 2).
2. We propose an estimator for the trimmed (thresholded) information density, named TIDE, and derive accompanying consistency and sample complexity guarantees. On the practical side, we present a neural network-based implementation for the TIDE (cf. Section 3).
3. We apply the obfuscation mechanism in Section 2 for image obfuscation (McPherson et al., 2016; Oh et al., 2017; Wu et al., 2018) with GENKI-4k (MPLab, 2009) and Celebrity Attributes (CelebA) (Liu et al., 2015) datasets, and for identifying politically-charged terms in Tweets collected from online media (Rachez, 2017) (cf. Section 4). These experiments provide evidence that the TIDE can potentially serve as a building block in the design of obfuscation mechanisms.

Proofs, experimental details, discussions, and additional experiments on synthetic data are provided in the Supplement. Source code for reproducing our experimental results is given at Hsu (2020).

Related Work The problem of balancing the competing objectives of providing meaningful information and inference from disclosed data, on the one hand, and obfuscating sensitive information, on the other hand, has been widely studied in information-theoretic privacy (cf., e.g., Calmon and Fawaz (2012); Issa et al. (2018); Diaz et al. (2018)). Following the information-theoretic trend, these works exploit average measures (in particular mutual information and its variants) to obfuscate data. Recently, information obfuscation has been achieved using neural networks. For example, in Bertran et al. (2019), an optimization problem similar to the *privacy funnel* (Makhdoumi et al., 2014) is formulated to train a neural network to automatically obfuscate sensitive information while maintaining utility. In Chen et al. (2019); Huang et al. (2017), neural generative models are introduced to generate "privatized" data that resemble the original data. These works rely on neural networks to select and perturb features. The approach proposed here is different in the sense that it first identifies the information-leaking features using the information density, and then applies local obfuscation only on these features.

The two-stepped approach of first identifying the information-leaking features and then perturbing those features is inspired by the instance-based additive mechanism of Nissim et al. (2007) in the DP setting. In fact, the information density appears in DP under the name of *privacy loss* variable (cf., e.g., Dwork and Rothblum (2016), thereby connecting DP and information-theoretic quantities, e.g., mutual information DP (Cuff and Yu, 2016) and Rényi DP (Mironov, 2017). Despite this connection, we emphasize that our approach is fundamentally different from DP, in that we consider prior distribution on sensitive attributes and also we allow correlation among features (see, e.g., (Kifer and Machanavajjhala, 2011) for the limitations of DP for correlated data).

Estimating information density from samples is connected to density ratio estimation (Nguyen et al., 2010; Liu et al., 2017; Yamada et al., 2011) — a fundamental task in various applications of machine learning and statistics, including outlier detection (Smola et al., 2009), transfer learning (Sugiyama et al., 2007), and generative adversarial networks (Goodfellow et al., 2014). A naïve approach to determine the density ratio is to use the plug-in estimator, which is known to perform poorly (Vapnik, 2013) unless adequate parametric models (e.g., linear (Yamada et al., 2011), kernel (Sugiyama et al., 2012), or exponential family (Liu et al., 2017) models) are assumed. The two closest approaches to the trimmed information density estimation in this paper are (i) Nguyen et al. (2010), which proposed using the variational representation of f-divergences to convert information density estimation into an optimization problem over finite-complexity set of functions and (ii) Liu et al. (2017), which estimated the trimmed density ratio of variables from

exponential family distributions. We enforce a threshold on the information density when solving the optimization problem in the variational representation of f-divergences (see Section 3).

Notation Capital letters (e.g., X) denote random variables, and calligraphic letters (e.g., $\mathcal{X}$) denote sets. We denote the probability measure of $X \times S$ by $P_{X,S}$, the conditional probability measure of S given X by $P_{S|X}$, and the marginal probability measure of X and S by P_X and P_S respectively. We use $P_{S|X}(\cdot|x)$ and $P_{S|x}$ interchangeably. We represent the fact that X is distributed according to P_X by $X \sim P_X$. KL-divergence is given by $D_{\mathsf{KL}}(P_{S,X}\|P_S P_X) = \mathbb{E}_{P_{S,X}}[\log(P_{S,X}/P_S P_X)]$. We denote the realization (i.e., sample) drawn from a probability distribution by $x = (x_1, \cdots, x_j, \cdots, x_m)$, where x_j is the j^{th} feature for $j = 1, \cdots, m$. Similarly, X_j is the j^{th} feature of the data variable. We denote $[k] = [1, \cdots, k]$, $x^k = [x_1, \cdots, x_k]$, and $(z)_+ = \max\{z, 0\}$ for a scalar z. Finally, $\mathrm{I}_{d\times d}$ is the identity matrix of dimension d, and $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

2 Problem Formulation

We consider the setting where a user wishes to disclose data X (e.g., image, tweet) while controlling the information revealed about a (correlated) sensitive attribute S (e.g., emotion, political preference). The goal is to produce an obfuscated representation Y of X that discloses only negligible information about S. We assume that X consists of m features, i.e., $X = (X_1, \ldots, X_m)$, where each feature takes values in a compact set $\mathcal{X}$. Throughout this section, we assume that $(S, X) \sim P_{S,X}$ and the underlying distribution $P_{S,X}$ is given. This restrictive assumption will be dropped in the subsequent section.

One possible approach to obfuscate X is to independently perturb each feature (e.g., by adding noise to each pixel of an image). However, in many applications, only a few features of the data are correlated with the sensitive attribute, rendering adding independent noise highly sub-optimal. In this section, we propose an information-theoretic framework for data obfuscation which consists of two parts: First, we *identify information-leaking features*, and then obfuscate *only* those features. In particular, our framework allows the flexibility to obfuscate those features in accordance to privacy and utility requirements. This way, many features need not be perturbed, leading to a potential improvement in the utility of the disclosed data.

Our framework relies on an information-theoretic quantity called the *information density*, a term coined in Pinsker (1964) and has since been used in numerous applications in information theory and statistics, particularly in binary hypothesis testing (see, e.g., Neyman-Pearson Lemma (Cover and Thomas, 2012)).

Definition 1 (Information Density)**.** Given a pair of realization (s, x) of $(S, X) \sim P_{S,X}$, the information density between s and x is defined as

$$i(s;x) \triangleq \log \frac{P_{S,X}(s,x)}{P_S(s)P_X(x)} = \log \frac{P_{X|S}(x|s)}{P_X(x)}. \tag{1}$$

Similarly, information density can be defined for each feature x_j as

$$i(s;x_j) \triangleq \log \frac{P_{S,X_j}(s,x_j)}{P_S(s)P_{X_j}(x_j)}, \tag{2}$$

and the conditional information density between s and x_j given another feature x_r as

$$i(s;x_j|x_r) \triangleq \log \frac{P_{S,X_j|X_r}(s,x_j|x_r)}{P_{S|X_r}(s|x_r)P_{X_j|X_r}(x_j|x_r)}. \tag{3}$$

Intuitively, $i(s;x_j)$ evaluates the change of belief about s upon observing x_j. In particular, if $|i(s;x_j)|$ is small, then x_j does not significantly contribute in increasing the belief of an adversary about s, since $P_{S|X}(s|x_j) \approx P_S(s)$. This, however, does not mean that x_j can be disclosed "as is" without incurring an information leakage risk. To see why, consider, for example, that $m = 2$, X_1 and X_2 are independent and uniform binary random variables, and $S = X_1 + X_2$ (modulo 2). Although $i(s;x_1) = i(s;x_2) = 0$ for any realization (s, x_1, x_2) of (S, X_1, X_2), the release of both x_1 and x_2 would allow perfect reconstruction of s. To account for such inferences of sensitive attributes, we consider the conditional information density as a yardstick for identifying information-leaking features.

Definition 2 (Information-Leaking Feature)**.** Given an observed sample $x = (x_1, \cdots, x_m)$, $j \in [m]$, and $\varepsilon \geq 0$, the feature x_j is said to be an *ε-information-leaking feature* if there exists a sensitive attribute s such that $|i(s;x_j|x^{j-1})| > \varepsilon$.

The threshold ε is a tradeoff parameter between information leakage risk and the utility of the disclosed data (e.g., the quality of an image). Notice that if the data is not equipped with a natural ordering (e.g., time series), we can choose an arbitrary ordering for the conditioning features x^{j-1} (cf. Section 4.1 for an example in images).

2.1 A Naïve Obfuscation Mechanism

Given any $j \in [m]$, $\varepsilon \geq 0$, and all features x^{j-1}, define

$$\begin{aligned} &B_j^{\varepsilon}(x^{j-1}) \\ &\quad \triangleq \{x \in \mathcal{X} : |i(s;x|x^{j-1})| > \varepsilon \text{ for some } s \in \mathcal{S}\}. \end{aligned} \tag{4}$$

If $x_j \notin B_j^{\varepsilon}(x^{j-1})$, then it can be disclosed "as is" because it cannot be used to infer sensitive attributes

given all the previous features. On the other hand, each feature $x_j \in B_j^\varepsilon(x^{j-1})$ is required to be obfuscated. To do so, we shall pass all such features sequentially through an *obfuscation mechanisms* to ensure that they no longer belong to $B_j^\varepsilon(x^{j-1})$.

Consider the mechanisms $\mathcal{M}_j : \mathcal{X} \to \mathcal{X}$ such that if $x_j \notin B_j^\varepsilon(x^{j-1})$ then $\mathcal{M}_j(x_j) = x_j$ (deterministic) and if $x_j \in B_j^\varepsilon(x^{j-1})$ then $\mathcal{M}_j(x_j)$ generates Y_j a random variable from a distribution to be designed. A natural question raised here is: how should information obfuscation be measured? To answer this question, we introduce the *obfuscation metric* $\Pr(|i(s;Y_j|y^{j-1}|) > \varepsilon)$ and require

$$\Pr(|i(s;Y_j|y^{j-1}|) > \varepsilon) \leq \frac{\delta}{m}, \tag{5}$$

for all $s \in \mathcal{S}$, where y^{j-1} is any output of the $\mathcal{M}_1(x_1), \ldots, \mathcal{M}_{j-1}(x_{j-1})$. Although this metric is intuitive, it presents a serious drawback for use in practice. Any reasonable mechanism must be immune to post-processing: any processing of the mechanism's output should only decrease the information leakage risk or equivalently the obfuscation metric. However, the obfuscation metric in (5) may violate this property. To see this, let $m = 1$ and $\widetilde{Y}$ be obtained by applying an arbitrary post-processing to Y the output of the mechanism $\mathcal{M}_1$ satisfying the obfuscation metric $\Pr(i(Y;s) > \varepsilon) \leq \delta$ for all s. Immunity to post-processing is then equivalent to requiring

$$\Pr(i(s;\widetilde{Y}) > \varepsilon) \leq \Pr(i(s;Y) > \varepsilon), \tag{6}$$

for all s, $\varepsilon \geq 0$ and $\delta \in [0,1]$. However, we show in the following that there must exist some ε for which (6) is violated. To see this, notice that $\mathbb{E}\left[\frac{P_{\widetilde{Y}|S}(\widetilde{Y}|s)}{P_{\widetilde{Y}}(\widetilde{Y})}\right] = \mathbb{E}\left[\frac{P_{Y|S}(Y|s)}{P_Y(Y)}\right] = 1$ and hence we have

$$\int_0^\infty \Pr(e^{i(s;\widetilde{Y})} \geq t)\mathrm{d}t = \int_0^\infty \Pr(e^{i(s;Y)} \geq t)\mathrm{d}t. \tag{7}$$

Suppose Eq. (6) is equality for all ε except an ε_0 for which it holds with strict inequality, then the integrals in (7) cannot be equal; therefore, Eq. (6) must hold with equality for all $\varepsilon \geq 0$ which in turn implies

$$D_{\mathsf{KL}}(P_{\widetilde{Y}|s}\|P_{\widetilde{Y}}) = D_{\mathsf{KL}}(P_{Y|s}\|P_Y). \tag{8}$$

However, according to data processing inequality for KL divergence, Eq. (8) cannot hold true in general. Therefore, there must exist some ε for which (6) does not hold. For more details about this construction, see Liu et al. (2017).

Next, we propose another metric in terms of a certain f-divergence, the so-called E_γ-divergence, and show that it implies (5) while being immune to post-processing.

2.2 E_γ-Divergence

To address the issue raised above, we resort to a particular divergence metric between two probability distributions called E_γ-divergence, and show that this divergence bounds an appropriately weighted tail distributions of $i(s;Y)$.

Definition 3 (E_γ-Divergence (Polyanskiy et al., 2010))**.** Given two probability distributions P and Q defined on the same support set $\mathcal{A}$ and $\gamma \geq 1$, we define E_γ-divergence as

$$\mathsf{E}_\gamma(P\|Q) \triangleq \sup_{A\subset\mathcal{A}} P(A) - \gamma Q(A) \tag{9}$$

$$= \int_{a\in\mathcal{A}} (\mathrm{d}P(a) - \gamma \mathrm{d}Q(a))_+, \tag{10}$$

where the equality comes from the fact that the optimizer in (9) is $\mathcal{A}^* = \{a \in \mathcal{A} | P(a) - \gamma Q(a) \geq 0\}$.

E_γ-divergence has been considered in various fields; for example, it appears in DP literature as an equivalent definition for differentially private mechanisms (see e.g, Barthe and Olmedo (2013); Balle et al. (2019)), in statistics as the probability of correct decision in Bayesian binary hypothesis testing (Polyanskiy et al., 2010), and in information theory for proving general channel coding converse results (Polyanskiy et al., 2010; Polyanskiy and Verdú, 2010).

Notice that $\mathsf{E}_\gamma(P\|Q) \leq 1$ for all $\gamma \geq 1$ and any pair of distributions (P,Q). It is clear that the constraint $\mathsf{E}_\gamma(P_Y\|P_{Y|s}) \leq \delta$ for some $\delta \in (0,1)$ ensures that $P_Y(A) - \gamma P_{Y|s}(A) \leq \delta$ for *all* subsets $A \subset \mathcal{X}$ and in particular $P_Y(\mathcal{A}^*) \leq \delta$. Since for $\gamma = e^\varepsilon$, the set $\mathcal{A}^*$ corresponds to the tail events of the random variable $i(Y;s)$, we henceforth assume $\gamma = e^\varepsilon$. Note also that to have control on both tail events $\{i(Y;s) < -\varepsilon\}$ and $\{i(Y;s) > \varepsilon\}$, we need to consider both $\mathsf{E}_{e^\varepsilon}(P_Y\|P_{Y|s}) \leq \delta$ and $\mathsf{E}_{e^\varepsilon}(P_{Y|s}\|P_Y) \leq \delta$. In the sequel, we present our results only for $\mathsf{E}_{e^\varepsilon}(P_{Y|s}\|P_Y) \leq \delta$. The results for the reversed divergence can be derived *mutatis mutandis*.

Having this divergence at our disposal, we can now propose obfuscation criteria for the mechanisms $\{\mathcal{M}_j\}$. As before, if $x_j \notin B_j^\varepsilon(x^{j-1})$, we set $\mathcal{M}_j(x_j) = x_j$; otherwise, we shall construct randomized mechanism $\mathcal{M}_j : \mathcal{X} \to \mathcal{X}$ such that $\mathcal{M}_j(x_j) = Y_j$ satisfies

$$\mathsf{E}_{e^\varepsilon}(P_{Y_j|s,y^{j-1}}\|P_{Y_j|y^{j-1}}) \leq \frac{\delta}{m}, \tag{11}$$

where y^{j-1} is a realizations of all previous mechanisms $\mathcal{M}_1(x_1), \ldots, \mathcal{M}_{j-1}(x_{j-1})$. The factor $\frac{1}{m}$ in the right-hand side of (11) is only for the sake of normalization (to be clarified in Theorem 2).

It is clear from (9) that upper bounds on $\mathsf{E}_\gamma(P_{Y_j|s,y^{j-1}}\|P_{Y_j|y^{j-1}})$ directly translate into low-

leakage guarantee (5). Furthermore, since E_γ-divergence belongs to the family of f-divergences (Sason and Verdú, 2016), it satisfies the data processing inequality which in turn implies that mechanisms satisfying (11) are immune to post-processing.

To even further justify the choice of E_γ-divergence as a "proxy" for the obfuscation metric in (5), we prove in the following theorem an equivalent formula for $\mathsf{E}_{e^\varepsilon}(P_{Y_j|s,y^{j-1}}\|P_{Y_j|y^{j-1}})$ in terms of the tail distribution $\Pr(i(s;Y_j|y^{j-1})>t)$ for $t\geq 0$.

Theorem 1 (Tail Distribution Formula)**.** Given distributions $P_{Y_j|s,y^{j-1}}$ and $P_{Y_j|y^{j-1}}$, we have

$$\mathsf{E}_{e^\varepsilon}(P_{Y_j|s,y^{j-1}}\|P_{Y_j|y^{j-1}}) = e^\varepsilon\int_\varepsilon^\infty e^{-t}\Pr(i(s;Y_j|y^{j-1})>t)\mathrm{d}t. \tag{12}$$

This result provides an operational interpretation for E_γ-divergence for our obfuscation setting. More precisely, $\mathsf{E}_{e^\varepsilon}(P_{Y_j|s,y^{j-1}}\|P_{Y_j|y^{j-1}})\leq\delta$ enforces the events $\{i(Y;s)>t\}$ to have small aggregate (weighted) probability for *all* $t\geq\varepsilon$.

Next, we address the composition property of the above mechanisms: If each mechanism $\mathcal{M}_j$ satisfies (11), then so does the composed mechanism $\mathcal{M}=(\mathcal{M}_1,\ldots,\mathcal{M}_m)$ with parameters $m\varepsilon$ and δ. Recall that $Y=(Y_1,\ldots,Y_m)$ is the output of the mechanism $\mathcal{M}$.

Theorem 2 (Composition)**.** For all mechanisms $\mathcal{M}_j, j\in[m]$ satisfying (11), we have for all $s\in\mathcal{S}$

$$\mathsf{E}_{e^{m\varepsilon}}(P_{Y|s}\|P_Y)\leq\delta. \tag{13}$$

This theorem states that a guarantee for each feature, given by (11), will result in a meaningful guarantee for the whole sample. This, in particular, demonstrates the need for conditional information density in Definition 2, as opposed to the unconditional one.

2.3 A Gaussian Obfuscation Mechanism

We next give an explicit construction of mechanisms $\{\mathcal{M}_j\}$ satisfying (11). Here, we assume that each feature $x_j\in C$ where C is a compact subset of $\mathbb{R}^r$. Recall that each mechanism $\mathcal{M}_j$ is required to generate Y_j satisfying (11). As a simple approach to enforce this guarantee, we propose the additive Gaussian mechanism; that is, for each given $j\in[m]$, ε, and $x^{j-1}\in\mathcal{X}^{j-1}$, we consider the following mechanism

$$Y_j = x_j + \lambda\mathbf{1}_{\{x_j\in B_j^\varepsilon(x^{j-1})\}}N, \tag{14}$$

where N is an independent standard Gaussian noise $\mathcal{N}(0,\mathrm{I}_{r\times r})$ and $\lambda>0$ is determined according to the following theorem.

Theorem 3 (Gaussian Obfuscation)**.** The Gaussian obfuscation mechanism (14) satisfies (11) if λ satisfies

$$\theta_{e^\varepsilon}(K,\lambda)\leq\frac{\delta}{m}, \tag{15}$$

where K is the radius of C, i.e., $K=\max_{w\in C}\|w\|$, and for any $a>0$

$$\theta_{e^\varepsilon}(a,\lambda)\triangleq\mathsf{Q}\left(\frac{\lambda\varepsilon}{a}-\frac{a}{2\lambda}\right)-e^\varepsilon\mathsf{Q}\left(\frac{\lambda\varepsilon}{a}+\frac{a}{2\lambda}\right), \tag{16}$$

where $\mathsf{Q}(v)=\Pr(\mathcal{N}(0,1)\geq v)=\int_v^\infty\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\mathrm{d}t$.

In light of this theorem, if $\varepsilon\approx 0$, then the noise variance λ must be of order $O(\frac{K}{-\log(1-\frac{\delta}{m})})$. The exact value of noise variance, however, cannot be derived as there is no analytic expression for the Q function.

We have thus far made the information-theoretic assumption that the underlying distribution $P_{S,X}$ is given and, consequently, the information density is known exactly. In the following section, we propose a *data-driven* estimator for information density which renders our proposed mechanism applicable to real-world datasets.

3 Trimmed Information Density

The obfuscation mechanism in Section 2 relies on the conditional information density $i(s;x_j|x^{j-1})$ to identify the set of information-leaking features. Notice that, since information density satisfies the chain rule, i.e.

$$i(s;x_j|x^{j-1}) = i(s;x^j) - i(s;x^{j-1}), \tag{17}$$

an estimate of $i(s;x_j|x^{j-1})$ can be constructed by estimates of $i(s;x^j)$ and $i(s;x^{j-1})$.

In general, exact estimation of the information density is hard due to its unboundeness. However, we do not need the exact estimation; instead, we only need to know if the absolute value of the conditional information density is larger than the threshold ε (Definition 2). In other words, estimating the *trimmed* information density is sufficient for obfuscation purposes. Moreover, the tail of the information density satisfies (Polyanskiy et al., 2010)

$$\Pr\left\{i(s;X_j)>t\right\}\leq e^{-t},\ \forall s, \tag{18}$$

indicating that the estimation error caused by trimming the information densities can be controlled. Exploiting the property in (18), in this section, we propose a consistent and scalable estimator for the *trimmed information density*, called the TIDE, and show that estimating the trimmed information density can be easier than estimating the exact information density in terms of sample complexity.

3.1 Trimmed Information Density Estimator

TIDE is based on a variational representation of KL divergence[1] known as the Donsker-Varadhan (DV) representation (Donsker and Varadhan, 1983), given by

$$D_{\mathsf{KL}}(P_{S,X}\|P_S P_X) = \sup_{g:\mathcal{S}\times\mathcal{X}\to\mathbb{R}} \{\mathbb{E}_{P_{S,X}}[g(S,X)] - \log \mathbb{E}_{P_S P_X}[e^{g(S,X)}]\}. \quad (19)$$

Recall that $D_{\mathsf{KL}}(P_{S,X}\|P_S P_X)$ is equal to the mutual information $I(S;X)$ between S and X, which is in fact the expected information density $\mathbb{E}_{P_{S,X}}[i(S,X)]$. It can be shown that the maximizer g^* of (19) is exactly the information density, i.e., $g^*(s,x) = i(s;x)$. Hence, the problem of estimating information density is equivalent to solving the functional optimization problem (19) given access to samples drawn from $P_{S,X}$.

Since the search space in (19) is unconstrained, directly solving the optimization by computing the empirical expectations would fail in general. One practical approach is to restrict the search space to a family $\mathcal{G}(\Theta)$ of continuous and bounded (by M) functions g_θ parameterized by θ in a compact domain $\Theta \subset \mathbb{R}^d$, where d is the number of parameters. The new constrained optimization problem corresponds to approximating the information density by a bounded function, thus the name *trimmed* information density.

The TIDE is then given by

$$\hat{g}_n \triangleq \operatorname*{argmax}_{g_\theta\in\mathcal{G}(\Theta)} \{\mathbb{E}_{P_{S_n,X_n}}[g_\theta(S,X)] - \log \mathbb{E}_{P_{S_n}P_{X_n}}[e^{g_\theta(S,X)}]\}, \quad (20)$$

where P_{S_n,X_n} and $P_{S_n}P_{X_n}$ denote the empirical distributions of $P_{S,X}$ and $P_S P_X$ by n samples, respectively.

3.2 Consistency and Sample Complexity

The TIDE obtained by solving (20) belongs to a broader class of *extremum estimators* (Amemiya, 1985) of the form $\hat{a} = \operatorname{argmax}_{a\in\mathcal{A}} \Lambda_n(a)$, where $\Lambda_n(a)$ is an objective function and $\mathcal{A}$ is a parameter space. The consistency of extremum estimators is guaranteed by the Newey-McFadden Lemma (Newey and McFadden, 1994) (cf. Supplement), which in turn implies the consistency of the TIDE, as stated in the following theorem.

Theorem 4 (Consistency)**.** If $\mathcal{G}(\Theta)$ is the family of continuous and bounded functions (with large enough M) parameterized by θ taking values in a compact domain Θ, then the TIDE (20) is consistent, i.e., for any $\eta > 0$, there exist $N > 0$ such that for all $n > N$, we have $|\hat{g}_n(s,x) - g^*(s,x)| \leq \eta$ with high probability for all $s \in \mathcal{S}$ and $x \in \mathcal{X}$.

We turn our attention to deriving the sample complexity of the TIDE. We make further assumption that functions in $\mathcal{G}(\Theta)$ are Lipschitz, and use (18) to prove the following theorem. To avoid technical complications, we assume that $\mathbb{E}_{P_{S,X}}[g(S,X)]$ and $\mathbb{E}_{P_S P_X}[e^{g(S,X)}]$ are finite for all functions g in $\mathcal{G}(\Theta)$.

Theorem 5 (Sample Complexity)**.** Assume that functions in $\mathcal{G}(\Theta)$ are bounded by M and Lipschitz with respect to θ, and $\Theta \subset \mathbb{R}^d$ is compact. Then we have $|\hat{g}_n(s,x) - g^*(s,x)| \leq \eta$ with probability at least $1 - e^{-M}$, for all $s \in \mathcal{S}$ and $x \in \mathcal{X}$, where $n = O(\frac{M^3 d}{\eta^2})$.

Observe that trimming the information density is crucial for the bound in the previous theorem to hold: if $M \to \infty$ (i.e., estimating the exact information density), the sample complexity of the TIDE grows to infinity and the result is vacuous. In fact, we need to restrict the search space to all continuous and bounded functions $\mathcal{G}$ to exactly approximate the trimmed information density. However, for computational reason, we assume that these functions can be parameterized by a compact domain Θ, and the complexity of the family $\mathcal{G}(\Theta)$ is characterized by its number of parameters d. As the complexity of the functions $d \to \infty$, meaning the search space is too large, the sample complexity goes to infinty. This assumption allows us to approximate the functions in $\mathcal{G}(\Theta)$ by neural networks, where Θ is the weights in all layers, as we will see next.

3.3 Implementation

In practice, we use the set of functions representable by a neural network with output clipped to $[-M, M]$ to approximate the set of continuous and bounded functions $g(s,x)$ in $\mathcal{G}$. By sampling (s,x) from $P_{S,X}$ and from $P_S \times P_X$ for the first and second expectations in (20), we can fit the neural network. After training, the $g(s,x)$ outputs the estimate of the trimmed information density of samples $|i(s;x)| \leq M$. In order to reconstruct the conditional information density by the chain rule (17), we compute $g(s,x^j)$ for $i(s,x^j)$ and $g(s,x^{j-1})$ for $i(s,x^{j-1})$; then the $i(s,x^j) - i(s,x^{j-1})$ gives the desired conditional information density $|i(s;x_j|x^{j-1})| \leq 2M$.

4 Experiments

The experiments contain two parts. First, we investigate image obfuscation (McPherson et al., 2016; Oh et al., 2017; Wu et al., 2018) as a common use case of our approach with the GENKI-4k (MPLab, 2009) and Celeberity Attributes (Liu et al., 2015) datasets. Second, we demonstrate how TIDE can be possibly used to discover politically-charged terms in the Tweets of online media (Rachez, 2017). Detailed experimental

[1] Other f-divergences (Csiszár, 1967; Sason and Verdú, 2016) could also be used, see the Supplement for more details.

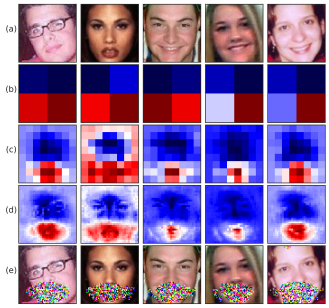

Figure 1: Row (a) shows original images. Rows (b), (c) and (d) show the information-leaking patches found by the TIDE (20) with patch sizes 32×32, 8×8 and 2×2 pixels respectively (color red indicates higher value). Row (e) shows the Gaussian obfuscation mechanism (14) on row (d) with $\varepsilon = 0.5$ and $\lambda = 1.0$, which successfully hide the sensitive attribute of emotion. The information-leaking patches is easy to interpret: the TIDE focuses more on the mouth area as the patches become finer.

setups (e.g., architecture of the function g in TIDE, training details) and additional experiments on Gaussian synthetic data are provided in the Supplement.

4.1 Image Obfuscation

A common application of information obfuscation is image obfuscation, where we aim to hide information related to a given sensitive attribute in an image. Unlike existing works which rely on neural networks to select and perturb features (McPherson et al., 2016), we apply the TIDE to identify information-leaking features for the Gaussian obfuscation mechanism (Section 2). We split x into a grid, where each "patch" of size $p \times p$ pixels in the grid represents the low-level features x_j of the image x. It is a common method to extract low-level features in an image. We number each x_j in an image from the upper-left corner to the lower-right, and use the TIDE (with $M = 3$) to determine the information-leaking features by (4), and demonstrate our obfuscation approach on two datasets: the GENKI-4k and Celeberity Attributes datasets.

4.1.1 GENKI-4K Dataset

This dataset contains 2400 images for training and 600 for testing, where each image x is a 64×64 pixels face that has emotion smiling ($s = 1$) or not ($s = 0$). We select 10 faces for illustration in Figure 1. When the patch size is 32×32 (4 patches), the TIDE flags the lower two patches to be information-leaking. As the patch becomes finer, the information-leaking patches concentrate to the mouse area; thus when applying the Gaussian obfuscation mechanism, it is visually possible to identify the gender of the subject but with their emotion obfuscated. The leakage guarantee in Theorem 3, $\delta/m \approx 0.24$, can be computed by (16) with $\varepsilon = 0.5$ and $K = 1$ since the images are normalized. The TIDE not only reveals the patches informative of emotion, but also captures the contour of faces.

We train an adversary that can classify the emotion of the subject with accuracy 92.04%, and report the classification accuracy of the Gaussian obfuscation mechanism ($\lambda = 1$ in (14)) under different patch sizes and threshold ε in Table 1. When $\varepsilon = \infty$ (i.e. $B_j^{\varepsilon}(x^{j-1}) = \phi$ for all j), no patch is identified by the TIDE, and therefore the performances are the same as the adversary. A simple mechanism to hide the emotion in images is adding Gaussian noise onto the Lower Half of the Image (LHI). As a comparison, the results of LHI and random guessing are also included in Table 1. The LHI gives similar performance when the patch size is 32×32 since when $\varepsilon = 0.5$, the lower two patches of the image will be identified as information-leaking for the mechanism (Figure 1 row (b)), but LHI will erase too much information that is not related to the emotion. The random guessing values correspond the to prior distribution of the emotion labels in the training set. Note that the information densities of the patches that leak privacy are around 0.6, according to Table 1. Hence, for $\varepsilon < 0.7$, all of the information-leaking patches around the mouth area are perturbed and the adversarial classifier has no useful information to infer emotion, resulting in a sharp drop of accuracy. In contrast, when ε is large, only a tiny portion of patches are perturbed and the classifier is still able to infer the emotion. This phenomenon can also be verified by the heatmaps in Figure 1 row (d), where only a small portion of patches have very high (crimson) values while most of the patches around the mouth have moderate (red/white) values. Moreover, when fixing an ε, the non-monotonicity of the accuracy in terms of the patch size is very data-driven since the mouth area may be located in different patches in the image.

Table 1: Classification accuracy of emotion obfuscation for the GENKI-4k dataset with different patch sizes $p \times p$ and threshold ε. Results on obfuscating the lower-half image by Gaussian noise (LHI) and on random guessing are shown as comparison.

	Classification Accuracy %				
$p \times p$ \ ε	0.5	0.6	0.7	0.8	∞
32×32	**50.54**	50.54	92.04	92.04	92.04
16×16	**50.72**	51.46	79.14	89.52	92.04
8×8	**50.93**	68.94	78.71	87.33	92.04
4×4	**50.60**	65.06	75.23	83.89	92.04
2×2	**50.64**	62.25	68.59	80.26	92.04
LHI	**50.58**	-	-	-	-
Guess	**50.41**	-	-	-	-

4.1.2 Celebrity Attributes (CelebA) Dataset

This more challenging dataset contains 202599 colorful high-resolution images, where each image is a 218×178-pixel face image of a celebrity with 40 distinct binary labels, including `smiling`, `gender`, `Arched Eyebrows`, etc. We select 100k images as X and the sensitive attribute S to be emotion as well for training the TIDE. In Figure 2, we randomly pick 4 images for illustration. Given a small patch size, the Gaussian obfuscation mechanism ($\lambda = 1$ in (14)) perturbs selective patches to hide the sensitive attribute while keeping other useful information (e.g. gender) intact.

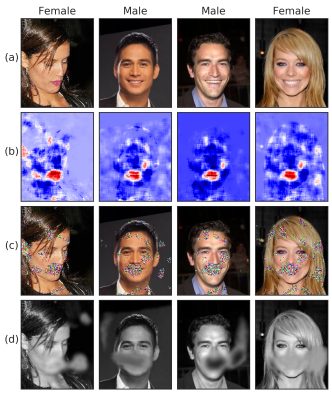

Figure 2: Row (a) shows original images. Row (b) shows the information-leaking patches with size 2×2 by the TIDE (color red indicates higher value). Row (c) shows the Gaussian obfuscation mechanism on row (b) with $\varepsilon = 0.74$ and $\lambda = 1$, and row (d) shows information obfuscation in Bertran et al. (2019) with the sensitive information budget equal to 0.72 bits.

Table 2: Comparison between our approach (with patch size 2×2) and Bertran et al. (2019) (ϵ here stands for the tolerance of sensitive information leak) on emotion and gender classification accuracy for the CelebA dataset.

Threshold	Classification Accuracy %			
	Our approach		Bertran et al. (2019)	
ε	Emotion	Gender	Emotion	Gender
∞	92.04	94.29	92.04	94.29
0.8	85.97	91.48	85.59	92.53
0.7	75.15	90.39	76.40	91.20
0.6	71.33	87.61	70.88	89.77
0.5	69.01	86.97	68.60	89.47
LHI	53.91	69.35	53.91	69.35
Guess	51.79	58.32	51.79	58.32

The leakage guarantee (Theorem 3), $\delta/m \approx 0.18$, can be computed by (16) with $\varepsilon = 0.74$ and $K = 1$. In Figure 2 row (d), we reproduce the method by Bertran et al. (2019) since it is the state-of-the-art result in information obfuscation and its implementation is publicly available. The main difference between our approach and theirs is that Bertran et al. (2019) requires an additionally pre-specified utility (i.e. the labels of gender), while our approach does not require such labels. As we can see, both methods shown in Figure 2 rows (c) and (d) obfuscate the mouth and some other area. We compute the number of perturbed pixels as a utility measure, i.e., the ℓ_0-norm of the difference between the original and obfuscated images, since the less number of perturbed pixels, the more information will be preserved. We evaluate the ratio of the number of perturbed pixels over the total number of pixels in percentage for the 4 images in Figure 2, and our method (from left to right) gives 10.12%, 9.05%, 7.88% and 15.91%, while (Bertran et al., 2019) gives 18.82%, 20.11%, 31.46% and 27.42%. Hence, our approach tends to obfuscate less of the subject's face.

We train two classifiers for emotion and gender, and report the accuracy of our approach and Bertran et al. (2019) in Table 2. Both methods block emotion recognition, effectively pushing the accuracy of the emotion classifier towards random guessing as ε decreases. More importantly, the gender classifier still performs well over the sanitized images. The experiments results in Table 2 shows that despite the theoretical flavor of our work and the clear level of interpretability, our method is comparable to the neural network-based algorithm in Bertran et al. (2019).

4.2 Information-Leaking Terms in Tweets

Finally, we showcase how the TIDE can be used in natural language to identify politically-charged terms in the Tweets from online media (Rachez, 2017). The information density is called the pointwise mutual information (PMI) in natural language processing to measure associations between words and labels (Church and Hanks, 1990). Since perturbation on languages is not yet well-defined (Alzantot et al., 2018), we do not perform the mechanism in (14), but focus on identifying information-leaking terms.

We collect $N = 75946$ Tweets from more than 20 online publishers (e.g. CNN, Bloomberg, New York Times), and determine their private attribute S as the political preference of being right-wing ($s = 0$) and left-wing ($s = 1$) according to Rachez (2017), where the numbers of samples with each political bias are equivalent. We pre-process the Tweets to keep only meaningful terms (i.e. pieces of words) and use bag-of-words representation (Manning et al., 2010) to tokenize all the pieces of words for each Tweet according to term frequency, ending up with 24657 terms (i.e. features x_j, $j \in [24657]$). We train the TIDE using the tokenized Tweets as x. In Figure 3, we show the estimate of trimmed conditional information density $i(s; x_j|x^{j-1})$ of each term. It is clear that some terms carry more information about the political bias. For instance, terms such as "Grand Old Party" and "National Rifle Association" associate with right-wing politics, and terms "Europe" and "liberal(s)" with the left. In this scenario, our approach could be eventually deployed as a plug-in to warn the users about potential political preference leaks before posting Tweets.

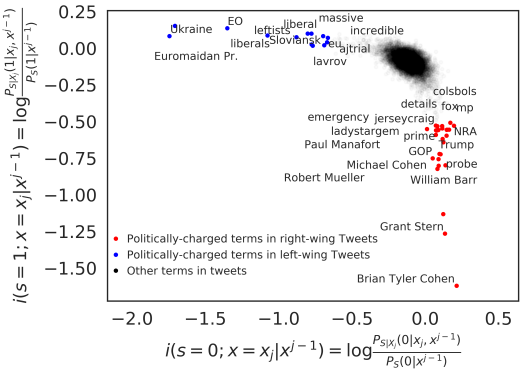

Figure 3: $i(s; x_j|x^{j-1})$ for terms in Tweets. GOP: Grand Old Party (i.e. the Republican Party), NRA: National Rifle Association, EO: Entrepreneurs' Organization, Euromaidan Pr.: Euromaidan Press.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. CIF-1845852 and CIF-1900750.

References

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., and Chang, K.-W. (2018). Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.

Amemiya, T. (1985). Asymptotic properties of extremum estimators. *Advanced econometrics, Harvard university press*.

Asoodeh, S., Diaz, M., Alajaji, F., and Linder, T. (2018). Estimation efficiency under privacy constraints. *IEEE Transactions on Information Theory*, 65(3):1512–1534.

Balle, B., Barthe, G., Gaboardi, M., and Geumlek, J. (2019). Privacy amplification by mixing and diffusion mechanisms. *ArXiv*, abs/1905.12264.

Balle, B. and Wang, Y.-X. (2018). Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *in Proc. of the International Conference on Machine Learning*.

Barthe, G. and Olmedo, F. (2013). Beyond differential privacy: Composition theorems and relational logic for f-divergences between probabilistic programs. In *Automata, Languages, and Programming*, pages 49–60, Berlin, Heidelberg. Springer Berlin Heidelberg.

Belghazi, I., Rajeswar, S., Baratin, A., Hjelm, R. D., and Courville, A. (2018). Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

Bertran, M., Martinez, N., Papadaki, A., Qiu, Q., Rodrigues, M., Reeves, G., and Sapiro, G. (2019). Adversarially learned representations for information obfuscation and inference. In *Proc. of International Conference on Machine Learning (ICML)*.

Bun, M. and Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography*, pages 635–658, Berlin, Heidelberg. Springer Berlin Heidelberg.

Calmon, F. P. and Fawaz, N. (2012). Privacy against statistical inference. In *Proc. of IEEE Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1401–1408.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109.

Chen, X., Navidi, T., Ermon, S., and Rajagopal, R. (2019). Distributed generation of privacy preserving data with user customization. *arXiv preprint arXiv:1904.09415*.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318.

Cuff, P. and Yu, L. (2016). Differential privacy as a mutual information constraint. In *Proc. of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

Diaz, M., Wang, H., Calmon, F. P., and Sankar, L. (2018). On the robustness of information-theoretic privacy measures and mechanisms. *arXiv preprint arXiv:1811.06057*.

Donsker, M. D. and Varadhan, S. S. (1983). Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212.

Dwork, C. and Rothblum, G. N. (2016). Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*.

Gao, W., Kannan, S., Oh, S., and Viswanath, P. (2017). Estimating mutual information for discrete-continuous mixtures. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, pages 5986–5997.

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

Hsu, H. (2020). Obfuscation via information density estimation. `https://github.com/HsiangHsu/ObduscationviaInformationDensityEstimation`.

Hsu, H., Asoodeh, S., Salamatian, S., and Calmon, F. P. (2018). Generalizing bottleneck problems. In *Proc. of IEEE International Symposium on Information Theory (ISIT)*.

Huang, C., Kairouz, P., Chen, X., Sankar, L., and Rajagopal, R. (2017). Context-aware generative adversarial privacy. *Entropy*, 19(12):656.

Issa, I., Wagner, A. B., and Kamath, S. (2018). An operational approach to information leakage. *arXiv preprint arXiv:1807.07878*.

Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 193–204, New York, NY, USA. ACM.

Liu, J., Cuff, P., and Verdú, S. (2017). E_γ-Resolvability. *IEEE Transactions on Information Theory*, 63(5):2629–2658.

Liu, S., Takeda, A., Suzuki, T., and Fukumizu, K. (2017). Trimmed density ratio estimation. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

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

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

Manning, C., Raghavan, P., and Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103.

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

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE.

MPLab, T. (2009). The MPLab GENKI Database, GENKI-4K Subset. `http://mplab.ucsd.edu`.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84. ACM.

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

Pinsker, M. S. (1964). *Information and information stability of random variables and processes*. San Francisco: Holden-Day.

Polyanskiy, Y., Poor, H. V., and Verdú, S. (2010). Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359.

Polyanskiy, Y. and Verdú, S. (2010). Arimoto channel coding converse and Rényi divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1327–1333.

Rachez, A. (2017). Predicting political bias with python. https://medium.com/linalgo/predict-political-bias-using-python-b8575eedef13. Accessed: 2019-03-21.

Sason, I. and Verdú, S. (2016). f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006.

Smola, A., Song, L., and Teo, C. H. (2009). Relative novelty detection. In *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research (JMLR)*, 8(May):985–1005.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Valiant, G. and Valiant, P. (2011). Estimating the unseen: An $n/\log n$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proc. of ACM symposium on Theory of computing (STOC)*.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Wu, Y. and Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720.

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

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M. (2011). Relative density-ratio estimation for robust distribution comparison. In *Proc. of Advances in neural information processing systems (NeurIPS)*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pages 325–333.