Learning Representations for Utility and Privacy: An Information-Theoretic Based Approach

Martin Bertran [1†]
martin.bertran@duke.edu

Natalia Martinez [1†*]
natalia.martinez@duke.edu

Afroditi Papadaki [2]
a.papadaki.17@ucl.ac.uk

Qiang Qiu [1]
qiuqiang@gmail.com

Miguel Rodrigues [2]
m.rodrigues@ucl.ac.uk

Guillermo Sapiro [1]
guillermo.sapiro@duke.edu

Abstract

Protecting a secret while disclosing related information (utility) is a well known challenge in privacy; both users and service providers can and should collaborate to protect privacy, and this paper addresses this paradigm. Here we analyze the limits of information-theoretic privacy, and use these to design a data-driven privacy-preserving representation of the disclosed data X that is maximally informative about the utility variable U and minimally informative about the secret variable S. We describe important use-case scenarios where the utility providers are willing to collaborate, at least partially, with the sanitization process. In this setting, we limit the possible sanitization functions to *space-preserving* transformations, where the same algorithm can be used to infer the utility variable on both sanitized and unsanitized data. We illustrate this approach though two use cases; *subject-within-subject*, where we tackle the problem of having an identity detector (from facial images) that works only on a consenting subset of users; and *emotion-and-gender*, where we tackle the issue of hiding independent variables, as is the case of hiding gender while preserving emotion detection.

1 Introduction, Challenges, and Contributions

We describe a scenario in which we have access to possibly high-dimensional data $X \in \mathcal{X}$, this data depends on two special latent variables U and S. U is called the utility latent variable, and is a variable we want to communicate, while S is called the secret, and is a variable we want to protect. We consider two agents, a service provider that wants to estimate U from X, and an adversary that wants to infer S from X. Our task is to communicate a sanitized representation of data X in such a way that a latent variable U can be inferred, but a sensitive latent variable S remains hidden.

The privacy learning algorithm is implemented as an adversarial game between the agent attempting to infer S and the privatizer, communicating US is a requirement of the privatizer. The learning algorithm is readily applicable to cases where the utility and secrecy variables U and S are either categorical, or where an assumption can be made on their distribution.

Contributions-

We describe a framework where we can apply information theoretic concepts of privacy even when the distributions on variables X, U and S are unknown. We derive an information-theoretic bound on privacy-preserving representations (mappings of X). The metrics induced by this bound are used to learn such a representation directly from data, without prior knowledge of the joint distribution of the observed data X and the latent variables U and S, but rather based on iid samples of these variables

Preprint. Work in progress.

drawn from the respective distirbutions. This process can accommodate for several user-specific privacy requirements, and can be modified to incorporate constraints about the service provider's existing utility inference algorithms.

Privacy-preserving representations learned with this framework have provable lower bounds, with a controllable parameter to control trade-off between the expected information leakage of the secret variable S and the expected information loss on the utility variable U.

We show example applications on facial images for two specific use cases; *subject-within-subject*, where we tackle the problem of having an identity detector (from facial images) that works only on a consenting subset of users; and *emotion-and-gender*, where we hide independent variables, as is the case of hiding gender while preserving emotion detection.

2 Information-theoretic bounds on privacy

Consider the utility and secret variables U and S defined over alphabets $\mathcal{U}$, $\mathcal{S}$,and the observed data variable X, defined over $\mathcal{X}$, with joint distribution $P_{X,U,S}$. Figure 1 illustrates this set-up, and shows the fundamental relationship of their entropies $H(\cdot)$ and mutual information between various variables.

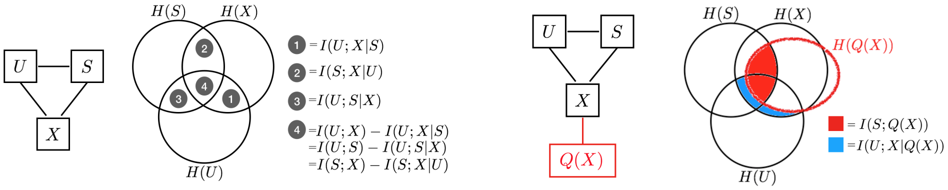


Figure 1: Left side shows the dependency structure of variables U, S, and X, along with important information measures. Right side extends this mapping to show the sanitized data $Q(X)$, conditionally independent of U and S given X. The information leakage $I(S;Q(X))$ and censured information $I(U;X \mid Q(X))$ shown in red and blue respectively cannot be simultaneously set to 0, since they are partially at odds.

We analyze the properties of any stochastic mapping $Q : \mathcal{X} \to \mathcal{Q}$, and measure the resulting mutual information between the transformed variable $Q(X)$ and our quantities of interest. Our goal is to find Q such that the information leakage from our sanitized data $I(S;Q(X))$ is minimized, while maximizing the shared information of the utility variable $I(U;Q(X))$, maximizing $I(U;Q(X))$ is equivalent to minimizing $I(U;X \mid Q(X))$. The quantity $I(U;X \mid Q(X))$ is the information X contains about U that is censured (lost) by the sanitization mapping Q.

Figure 1 illustrates $I(S;Q(X))$ and $I(U;X|Q(X))$. One can see that there exists a trade-off area, $I(U,S) - I(U,S|X)$, that is always included in the union of $I(S;Q(X))$ and $I(U;X|Q(X))$. The lower we make $I(S;Q(X))$, the higher we make the censored information $I(U;X|Q(X))$, and vice versa. This induces a lower bound over the performance of the best possible mappings $Q(X)$ that is formalized in the following lemma.

Lemma 1: Let X, U, S be three discrete random variables with joint probability distribution $P_{X,U,S}$. For any stochastic mapping $Q : \mathcal{X} \to \mathcal{Q}$ we have

$$[I(S;Q(X)] + [I(U;X|Q(X))] \geq I(U;S) - I(U;S|X). \tag{1}$$

Note that we can equivalently express

$$\begin{aligned} I(U;X \mid Q) &= E_{X,Q}\big[D_{KL}(p_{U|X} \mid\mid p_{U|Q})\big], \\ I(S,Q) &= E_{X,Q}\big[RD_{KL}(p_S \mid\mid p_{S|Q})\big], \end{aligned} \tag{2}$$

where D_{KL} and RD_{KL} are the Kullback-Leibler and reverse Kullback-Leibler divergence.

2.1 Defining a trainable loss metric

Assume that for any given stochastic transformation mapping $Q \sim Q(X)$, we have access to the posterior conditional probability distributions $P(S \mid Q)$, $P(U \mid Q)$, and $P(U \mid X)$. Assume we also have access to the prior distribution of $P(S)$. Inspired by the bounds from the previous section, the proposed privatizer loss is

$$min_Q(1-\alpha)E^2_{X,Q}[D_{KL}(p_{U|X} \mid\mid p_{U|Q})^2] + \alpha E^2_{X,Q}[RD_{KL}(p_S \mid\mid p_{S|Q})], \tag{3}$$

where $\alpha \in [0,1]$ is a tradeoff constant. A low α value implies a high degree of transparency (high utility), while a high value of α implies a high degree of privacy.

We can prove that for any $\alpha \in [0,1]$, and stochastic mapping $Q : \mathcal{X} \to \mathcal{Q}$ the solution Q^* to Eq.3 guarantees the following bounds,

$$\begin{aligned} E_{X,Q^*}[D_{KL}(p_{U|X} \mid\mid p_{U|Q^*})] &\geq \alpha[I(U,S) - I(U,S \mid X)], \\ E_{X,Q^*}[RD_{KL}(p_S \mid\mid p_{S|Q^*})] &\geq (1-\alpha)[I(U,S) - I(U,S \mid X)]. \end{aligned} \tag{4}$$

so, using the training objective in Eq.3, one is deliberately trying to match the left hand side of Eq 1 (which depends on Q) to the right hand side of eq 1 (which is a fundamental limit).

3 A Data-Driven implementation

The privatizer can attempt to minimize Eq.3 even when the joint distribution of $P(U,S,X)$ is not known by optimizing the following adversarial game:

$$\begin{aligned} \hat{\eta} &= \text{argmin}_\eta E_{X,S,Z}\big[-log(P_\eta(s|Q_{\hat{\theta}}(x,z))\big], \\ \hat{\psi} &= \text{argmin}_\psi E_{X,U,Z}\big[-log(P_\psi(u|Q_{\hat{\theta}}(x,z))\big], \\ \hat{\phi} &= \text{argmin}_\phi E_{X,U}\big[-log(P_\phi(u|x)\big], \\ \hat{\theta} &= \text{argmin}_\theta(1-\alpha)E^2_{X,U,Z}\big[D_{KL}(P_{\hat{\phi}}(u \mid x) \mid\mid P_{\hat{\psi}}(u \mid Q_{\hat{\theta}}(x,z)))]+ \\ &\quad + \alpha E^2_{X,S,Z}\big[RD_{KL}(P(s) \mid\mid P_{\hat{\eta}}(s \mid Q_{\hat{\theta}}(x,z)))]. \end{aligned} \tag{5}$$

Where $P_\eta(s|q)$, $P_\psi(u|q)$, and $P_\phi(u|x)$ estimators of the posterior of S and U after observing X and $Q_{\hat{\theta}}(x,z)$.

3.1 Privacy Under Fixed Utility Inference

A more interesting restriction arises when the utility inference algorithm $P_\phi(u|x)$ is given and cannot be modified. Furthermore, the privatizer is tasked with finding a mapping Q such that $P_\phi(u|x)$ applied to $Q_{\hat{\theta}}(x,z)$ still produces correct results, here we optimize

$$\begin{aligned} \hat{\eta} &= \text{argmin}_\eta E_{X,S,Z}\big[-log(P_\eta(s|Q_{\hat{\theta}}(x,z))\big], \\ \hat{\theta} &= \text{argmin}_\theta(1-\alpha)E^2_{X,U,Z}\big[D_{KL}(P_{\hat{\phi}}(u \mid x) \mid\mid P_{\hat{\phi}}(u \mid Q_{\hat{\theta}}(x,z)))]+ \\ &\quad + \alpha E^2_{X,S,Z}\big[RD_{KL}(P(s) \mid\mid P_{\hat{\eta}}(s \mid Q_{\hat{\theta}}(x,z)))]. \end{aligned} \tag{6}$$

4 Experiments and Results

The following examples are based on the framework presented in Figure 2. Here we have the three key agents mentioned before: (1) the utility algorithm that is used by the provider to estimate the information of interest. This algorithm can take the raw data (X) or the mapped data ($Q(X)$) and be able to infer the utility; (2) the secret algorithm that is able to operate on the raw data and the mapped data to infer the secret; (3) the privatizer that learns a space preserving mapping Q that allows the provider to learn the utility but prevents the secret algorithm to infer the secret.

4.1 Subject Within Subject

We analyze the *subject-within-subject* problem. Here, we want that a pretrained face verification algorithm can only verify the identity of a consenting subset of users, while blocking this task on nonconsenting users. We show this by training a space-preserving stochastic mapping Q on facial image data

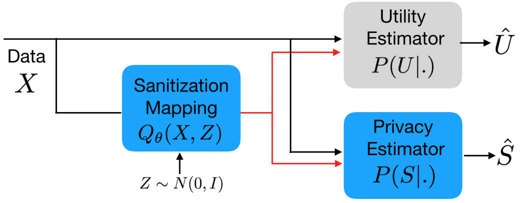
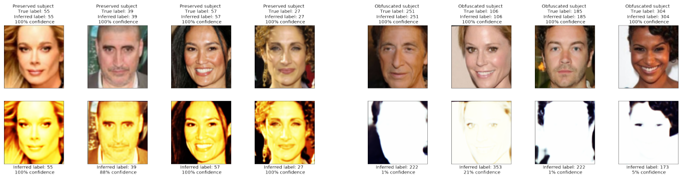


Figure 2: Three components of the collaborative privacy framework. Raw data can be directly fed into the secret and utility inferring algorithm. Since the privatization mapping is space preserving, the privatized data can also be directly fed to both tasks without any need for further adaptations.

X, where the utility variable U is a categorical variable over the consenting users, and the secrecy variable S is a categorical variable over the non-consenting users. We test this over the FaceScrub dataset [Kemelmacher-Shlizerman et al.(2016)Kemelmacher-Shlizerman, Seitz, Miller, and Brossard], using VGGFace2 [Cao et al.(2017)Cao, Shen, Xie, Parkhi, and Zisserman] as the base utility and secrecy inferring algorithm. The stochastic mapping was implemented using a stochastic adaptation of the UNET [Ronneberger et al.(2015)Ronneberger, Fischer, and Brox], where a Gaussian noise variable is learned along the image transformation, and is then injected before the upsampling stages.

Table 1 shows the top-5 categorical accuracy of the utility network over the sanitized data at various α points in the privacy-utility trade-off. Figure 3 show representantive images and samples of their sanitized counterparts.



(a) Filtered images of consenting users (CU)

(b) Filtered images of private users (PU)

	Top-5 Accuracy		
α	CU	PU	NU
0	99.8%	99.8%	99.8%
0.3	94.7%	86.1%	87.5%
0.35	93.5%	15.3%	33.6%
0.40	93.0%	14.9%	31.8%
0.45	89.4%	12.6%	30.2%
0.50	70.7%	10.9%	27.0%

Table 1: Subject detection on users

Figure 3: Left and center figures show images of consenting and nonconsenting (private) users respectively, along with their sanitized counterparts. The identity of consenting users is still easily verified, while the identity of nonconsenting users is effectively censored. Table on the right shows Top-5 accuracy performance of the subject detector after sanitization across several sanitation levels α. Performance is shown across 3 subsets, consenting users (CU), private users (PU), and new users (NU), this last group shows that the default behaviour of the sanitization function is to preserve privacy

We can see from Table 1 that the sanitization function is able to preserve information about the utility variable while effectively censoring the secret variable. This performance extends to unobserved images of the consenting subjects, and to images of new users.

4.2 Obfuscating Emotion While Preserving Gender

Here we continue to work on facial image data X, where utility variable U is gender, and the secret variable S is emotion (smiling/non-smiling). In this scenario, variables U and S are independent. We implement this over the CelebA dataset [Liu et al.(2015)Liu, Luo, Wang, and Tang], using Xception networks [Chollet(2017)] as our utility and privacy estimators. Table.2 shows the distribution of the utility and secrecy estimators over the sanitized data. Figure 4 shows example 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.

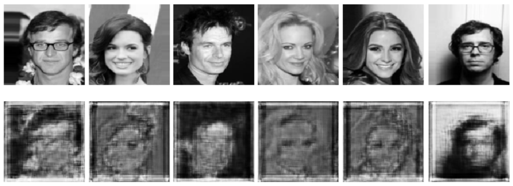


Figure 4: Images before and after sanitization for Gender (utility) vs Emotion (privacy).

Inference	Data	Male	Female
Gender	Raw	94.2%	94.5%
	Sanit.	91.2%	96.1%
Emotion	Raw	93.3%	92.0%
	Sanit.	54.0%	60.0%

Table 2: Gender and emotion detection on users on raw and sanitized data.

References

[Cao et al.(2017)Cao, Shen, Xie, Parkhi, and Zisserman] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2017.

[Chollet(2017)] François Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint*, pp. 1610–02357, 2017.

[Kemelmacher-Shlizerman et al.(2016)Kemelmacher-Shlizerman, Seitz, Miller, and Brossard] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. 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.

[Liu et al.(2015)Liu, Luo, Wang, and Tang] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[Ronneberger et al.(2015)Ronneberger, Fischer, and Brox] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.