Supplementary Material

Analysis of alternative loss functions

Let (X, U, S) be a collection of random vectors. we wish to optimize the following constrained optimization problem:

$$\min_{p(y|x)} I(U; X \mid Y) \quad s.t.\ I(S;Y) \leq k. \tag{11}$$

A natural approach, similar to the one used in the information bottleneck literature would be to minimize

$$\min_{p(y|x)} I(S;Y) - \beta I(U;Y), \tag{12}$$

where β controls the relative tradeoff between utility preservation and secret obfuscation. We show here how these problems are not equivalent when S and U are univariate gaussian and Y is a linear transformation:

$$X = \begin{bmatrix} U \\ S \end{bmatrix}; X \sim N(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}). \tag{13}$$

$$Y = OX + Z; Z \sim N(0,1); O \in \mathcal{R}_{1\times 2} \tag{14}$$

Without loss of generality we factorize Y as:

$$Y = U + R\rho S + \epsilon Z. \tag{15}$$

Under these assumptions the quantities of interest are:

$$\begin{aligned} I(U;Y) &= \frac{1}{2} log \frac{R^2\rho^2 + 2R\rho^2 + 1 + \epsilon^2}{R^2\rho^2(1-\rho^2) + \epsilon^2}, \\ I(S;Y) &= \frac{1}{2} log \frac{R^2\rho^2 + 2R\rho^2 + 1 + \epsilon^2}{(1-\rho^2) + \epsilon^2}, \end{aligned} \tag{16}$$

and their derivatives w.r.t R:

$$\begin{aligned} \frac{\partial I(U;Y)}{\partial R} &= \frac{\rho^2[R^2\rho^2(\rho^2-1) + R(\rho^2(1+\epsilon^2)-1) + \epsilon^2]}{[R^2\rho^2 + 2R\rho^2 + \epsilon^2 + 1][R^2\rho^2(1-\rho^2) + \epsilon^2]}, \\ \frac{\partial I(S;Y)}{\partial R} &= \frac{1-\rho^2+\epsilon^2}{[R^2\rho^2 + 2R\rho^2 + 1 + \epsilon^2]} \rho^2(R+1). \end{aligned} \tag{17}$$

From these equations we can conclude the following:

♣ $I(S;Y)$ has a minimum in $R = -1$ and is convex in R.

♣ $I(U;Y)$ has one minima, one maxima (R_ϵ^M) and an horizontal asymptote in $I(U;S)$ as $R \to \infty$.

♣ The local maxima of $I(U;Y)$ is attained at $R_\epsilon^M = \frac{-(1-\rho^2(1+\epsilon^2))+\sqrt{\Delta}}{2\rho^2(1-\rho^2)}$ where $\Delta = (\rho^2(1+\epsilon^2)-1)^2 + 4\rho^2(1-\rho^2)\epsilon^2$. When $\epsilon \to 0$ $R_\epsilon^M \to 0$.

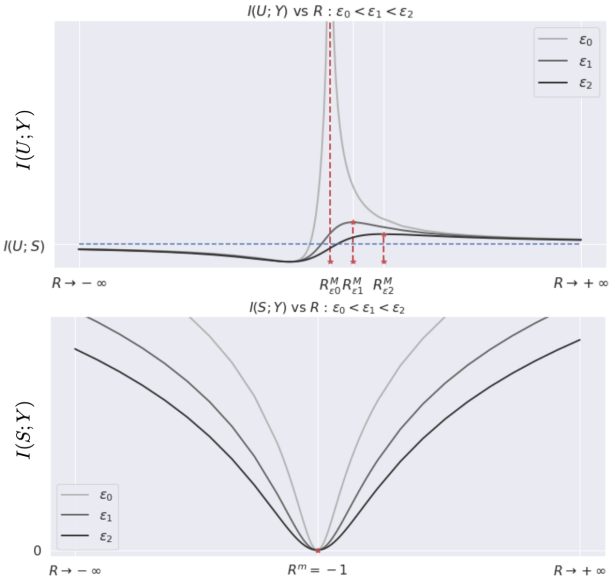

Figure 9. Top figure shows $I(U;Y)$ as a function of R for different ϵ values. Dotted red lines show their maximum values (R_ϵ^M). Bottom figure shows $I(S;Y)$, minimum value is reached when $R = -1$ independently of ϵ. Both figures are computed for $\rho = 0.5$

Figure 9 illustrates these points.

The solution to the constrained problem in Eq. 11 involves identifying the interval $[R_-^K, R_+^K]$ where $I(S;Y) \leq k$ and setting $R = R_\epsilon^M$ if $R_\epsilon^M \in [R_-^K, R_+^K]$ otherwise pick the interval extrema that maximizes $I(U;Y)$. Figure 10 shows the solution intervals as a function of k and the tradeoff curve that arises from varying k from 0 to ∞.

Solutions to unconstrained functional

We briefly analyze the behaviour of the unconstrained functional

$$L_\beta = I(S;Y) - \beta I(U;Y). \tag{18}$$

Since L_β is differentiable w.r.t. R and we do not impose any additional constraints, we analyze its derivatives to find the fixed points of the functional:

$$\begin{aligned} \frac{\partial L_\beta}{\partial R} &= \frac{1-\beta}{2} \frac{2R\rho^2 + 2\rho^2}{R^2\rho^2 + 2R\rho^2 + 1 + \epsilon^2} + \\ &\quad \frac{\beta}{2} \frac{2R\rho^2(1-\rho^2)}{R^2\rho^2(1-\rho^2) + \epsilon^2} \\ &\propto C^2\big[R^3 + R^2(1+\beta) + \\ &\quad R\frac{[(1-\beta)\epsilon^2 + \beta(1-\rho^2)(1+\epsilon^2)]}{\rho^2(1-\rho^2)} + \\ &\quad \frac{(1-\beta)\epsilon^2}{\rho^2(1-\rho^2)}\big]. \end{aligned}$$

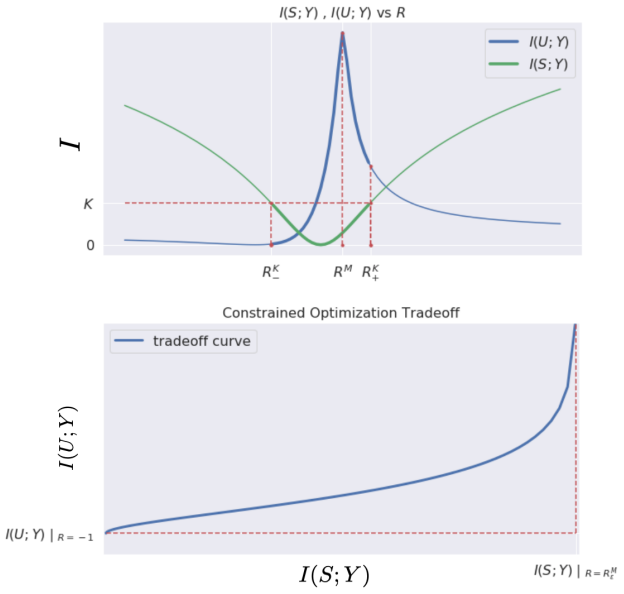

Figure 10. Top figure shows $I(S;Y)$ and $I(U;Y)$ as a function of R, feasible region is shown with a thicker lineweight. Bottom figure shows the optimal tradeoff curve we obtain by sweeping the constraint $K \in (0, \infty)$ in the $(I(S;Y), I(U;Y))$ space.

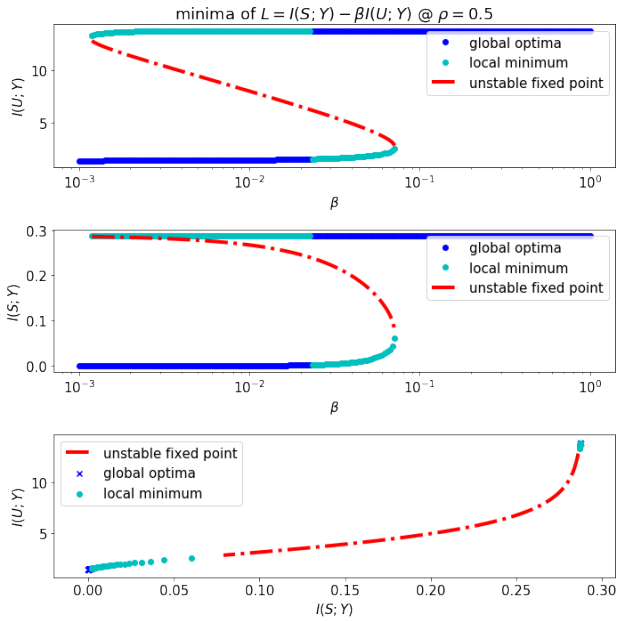

Figure 11. Fixed points of L_β as a function of β. Top figure shows attained values of $I(U;Y)$ across all fixed points for several β values. Middle figure shows $I(S;Y)$ under the same conditions. Bottom row shows the tradeoff curve attained by these fixed points. A large portion of non-trivial tradeoffs are obtained by either an unstable fixed point, or a local minima.

where C^2 is a function of R, ρ, β and ϵ, but is strictly positive.

From these equations we can conclude:

♣ This loss has at most 3 fixed points in R as a function of ϵ, β, ρ.

♣ All fixed points in R belong to the $[-1, 0]$ interval.

♣ For $\epsilon = 0$, $R = 0$ is always a stable fixed point.

♣ The second fixed point is unstable.

Figure 11 shows all fixed points of the unconstrained functional as a function of β. This figure also shows that this functional exhibits a behaviour similar to phase transition, where most tradeoff values are only attainable as unstable fixed points of the functional.

Remarks

Even in this simple case, the unconstrained functional is unable to produce arbitrary tradeoffs between $I(U;Y)$ and $I(S;Y)$. By contrast, the quadratic penalty method used throughout the text (Wright & Nocedal, 1999) is theoretically equivalent to the constrained optimization problem, and is empirically able to obtain arbitrary tradeoffs for this simplified example.

Lower Bound Estimation

The lower bound described in Lemma 2.1 requires solving the following constrained optimization problem:

$$\begin{aligned} \min_{p(y|u,s)} \; & I(U;X) - I(U;Y) \\ s.t. \; & I(S;Y) \leq k, \\ & I(U;Y) \leq I(U;X), \end{aligned} \tag{19}$$

where $p(y \mid u, s) : \mathcal{U} \times \mathcal{S} \to \mathcal{Y}$. There are two main barriers to overcome to compute this bound, the first is that we want an unconstrained formulation of this problem for ease of computation, and the second is that $\mathcal{Y}$ can potentially be very large, making the space of solutions $p(y \mid u, s)$ too large to efficiently optimize.

Fortunately, both issues can be circumvented efficiently by solving a sequence of small, unconstrained problems until a convergence criteria is met. Let $r > 1$ and $n_i = \lceil r \times n_{i-1} \rceil = |\mathcal{Y}_i|$ the alphabet size of the output variable of i-th problem (Y_i), where $n_0 = |\mathcal{U}|$. For each problem index i, we solve the following unconstrained problem:

$$\begin{aligned} \min_{p(y_i|u,s)} \; & I(U;X) - I(U;Y_i) + \lambda \max(I(S;Y_i) - k, 0)^2 \\ & + \lambda \max(I(U;Y_i) - I(U;X), 0)^2 \\ s.t. \; & |\mathcal{Y}_i| = \lceil n_i \rceil. \end{aligned} \tag{20}$$

Let B_i be the solution to Eq.(20). The procedure iterates the computation of B_i until $B_i \simeq B_{i-1}$. Computation of B_i is achieved through gradient descent as described in Algorithm 2.

In all the experiments shown in Section 5, bound computation took no more than 10 minutes and 3 iterations for $r = 1.5$.

Upper Bound Proof

We additionally show an extended proof of the equality stated Eq 7 in Lemma 2.2

Lemma 6.1. Let X, U, S be three discrete random variables with joint probability distribution $p(x, u, s)$. For any variable Y drawn

Algorithm 2 Restricted Cardinality Step

Input: Emprirical joint distribution of (U, S) $P_{u,s} \in \mathbf{R}_{|\mathcal{U}|\times|\mathcal{S}|}$; alphabet size n_i; hyperparameters $(lr, \lambda, k, I(U;X))$
Compute marginal distributions.
$P_u = \sum_j P_{u,j}$
$P_s = \sum_k P_{k,s}$
Compute expanded conditional distributions:
$P_{(u,s)|u} = \text{vec}\{P_{u,:} \oslash P_u\}$; $P_{(u,s)|u} \in \mathbf{R}_{(|\mathcal{U}|\times|\mathcal{S}|)\times|\mathcal{U}|}$
$P_{(u,s)|s} = \text{vec}\{P_{:,s} \oslash P_s\}$; $P_{(u,s)|s} \in \mathbf{R}_{(|\mathcal{U}|\times|\mathcal{S}|)\times|\mathcal{S}|}$
Initialize unnormalized transition matrix:
$D \in \mathbf{R}_{n_i\times(|\mathcal{U}|\times|\mathcal{S}|}$
repeat
  Normalize transition matrix:
  $P_{y|(u,s)} = \text{softmax}(D, \text{axis} = 0)$
  Compute marginal and conditional distributions of Y:
  $P_y = P_{y|(u,s)} \times \text{vec}\{P_{u,s}\}$
  $P_{y|u} = P_{y|(u,s)} \times P_{(u,s)|u}$
  $P_{y|s} = P_{y|(u,s)} \times P_{(u,s)|s}$
  Compute mutual informations:
  $I(Y;S) = \sum P_{y|s} \circ \ln P_{y|s} \oslash P_y$
  $I(Y;U) = \sum P_{y|u} \circ \ln P_{y|u} \oslash P_y$
  Compute functional value and do gradient descent:
  $L(D) = I(U;X) - I(Y;U) + \lambda \max(I(Y;S) - k, 0)^2 + \lambda \max(I(Y;U) - I(U;X), 0)^2$
  $D \leftarrow D - lr\nabla_D L(D)$
until Convergence
Return:
$P_{y|(u,s)}, L(D)$

from $p(y \mid x)$ and conditionally independent on U, S given X we have

$$\begin{aligned} &I(U;S) - I(U;S \mid X) \\ &= I(S;Y) - I(S;Y \mid U) + I(U;X \mid Y) - I(U;X \mid Y,S). \end{aligned} \tag{21}$$

Or, equivalently

$$\begin{aligned} &I(U;X \mid Y) + I(S;Y) \\ &= I(U;S) - I(U;S \mid X) + I(S;Y \mid U) + I(U;X \mid Y,S). \end{aligned} \tag{22}$$

Proof: From the Markov property we have

$$\begin{aligned} I(U;X \mid Y) &= I(U;X) - I(U;Y) \\ I(S;X \mid Y) &= I(S;X) - I(S;Y) \end{aligned} \tag{23}$$

By adding these two equations together, we obtain

$$\begin{aligned} &I(U;X \mid Y) + I(S;Y) \\ &= I(S;X) + I(U;X) - I(U;Y) - I(S;X \mid Y) \\ &= H(S) - H(S \mid X) - H(U \mid X) \\ &\quad + H(U \mid Y) - I(S;X \mid Y) \\ &= I(U;S) + H(S \mid U) - I(U;S \mid X) - H(U,S \mid X) \\ &\quad + H(U \mid Y) - I(S;X \mid Y) \\ &= I(U;S) - I(U;S \mid X) + H(S \mid U) - H(U,S \mid X) \\ &\quad + H(U \mid Y) - I(S;X \mid Y) \end{aligned} \tag{24}$$

Where we additionally used $I(U;S) + H(S \mid U) = H(S)$; $I(U;S \mid X) + H(U;S \mid X) = H(U \mid X) + H(S \mid X)$. The equality in Eq.(22) can be then proven by showing:

$$\begin{aligned} &H(S \mid U) - H(U,S \mid X) + H(U \mid Y) - I(S;X \mid Y) \\ &= H(S \mid U) - H(S \mid X) - H(U \mid S,X) \\ &\quad + H(U \mid Y) - H(S \mid Y) + H(S \mid X) \\ &= H(S \mid U) - H(U \mid S,X) + H(U \mid Y) - H(S \mid Y) \\ &= I(S;Y \mid U) + H(S \mid Y,U) - H(U \mid S,X) \\ &\quad + H(U \mid Y) - H(S \mid Y) \\ &= I(S;Y \mid U) + H(S \mid Y,U) + I(U;X \mid Y,S) \\ &\quad - H(U \mid Y,S) + H(U \mid Y) - H(S \mid Y) \\ &= I(S;Y \mid U) + I(U;X \mid Y,S) + H(S \mid Y,U) \\ &\quad - H(S \mid Y) + H(U \mid Y) - H(U \mid Y,S) \\ &= I(S;Y \mid U) + I(U;X \mid Y,S) - I(U;S \mid Y) \\ &\quad + I(U;S \mid Y) \\ &= I(S;Y \mid U) + I(U;X \mid Y,S) \end{aligned} \tag{25}$$

Where we used $H(S \mid U) = I(S;Y \mid U) + H(S \mid Y,U)$; $I(U;X \mid Y,S) = H(U \mid Y,S) - H(U \mid X,Y,S) = H(U \mid Y,S) - H(U \mid X,S)$. □

Domain-Preserving and Fixed Utility Inference Algorithm

Here we present the variant of Algorithm 1 described in Section 3.1, where we impose the additional constraint that the transformation must be a domain-preserving transformation, and the utility inference algorithm is given and cannot be modified. The algorithm is shown in Algorithm 3

Algorithm 3 Adversarial Information Obfuscation. Domain-Preserving and Fixed Utility

Input: data $\{(x_i, s_i, u_i)\}$; hyperparameters (lr, λ, k); utility inference algorithm $p_\phi(u \mid \cdot)$
$p(s)$ 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 sensitive inference networks:
  $H(\eta) = \frac{1}{b}\sum_{i=1}^{b} -logp_\eta(s_{(i)} \mid Q_\theta(x_{(i)}, z_{(i)}))$
  Stochastic gradient descent:
  $\eta \leftarrow \eta - lr\nabla_\eta H(\eta)$
  Evaluate unconstrained penalty loss:
  $\Theta(\theta) = \frac{1}{b}\sum_{i=1}^{b} log\frac{p_\phi(u_{(i)}|x_{(i)})}{p_\phi(u_{(i)}|Q_\theta(x_{(i)},z_{(i)}))} + \lambda\max(\frac{1}{b}\sum_{i=1}^{b} log\frac{p_\eta(s_{(i)}|Q_\theta(x_{(i)},z_{(i)}))}{p(s_{(i)})} - k, 0)^2$
  Stochastic gradient descent:
  $\theta \leftarrow \theta - lr\nabla_\theta\Theta(\theta)$
until Convergence

Implementation Details

We now describe the architectures, algorithms and hyperparameters used for each experiment.

Synthetic Data

In Section 4, we applied Algorithm 1 to synthetic data. Variables U and S were uniformly distributed on $|\mathcal{U}| = 6$ and $|\mathcal{S}| = 2$, the

observation process X is given by $X = (U, S)$. We controlled the joint distribution of $p(u, s)$ to obtain datasets with varying $I(U;S)$ values to do this, we used the following joint distribution: $p(u,s) \propto 1 + \beta\delta(u \bmod |\mathcal{S}| = s)$ where $\beta \geq 0$ directly impacts $I(U;S)$ ($\beta = 0$ yields $I(U;S) = 0$ while $\beta \to \infty$ yields $I(U;S) = \ln 2$, the maximum achievable value).

The posterior inference networks $p_\eta(S \mid \cdot), p_\psi(U \mid \cdot)$ and $p_\phi(U \mid \cdot)$ where implemented using the architecture shown in Figure 13 shows the detailed architecture. We tried two different filter architectures, a linear filter ($q_\theta(X, Z) = \Theta X + Z$, $Z \sim N(0, I)$), and an stochastic neural network whose architecture is also shown in Figure 13. Both filters were trained using Algorithm 1. Hyperparameters were chosen from the following set: learning rate $lr \in \{5e-5, 1e-4, 5e-4\}$, $\lambda \in \{1e2, 1e3, 1e4\}$, tolerance $k \in [0, I(U;S)]$. Figure 12 shows some of the learned representations Y on the linear and nonlinear architectures, and using the RCS sequence algorithm 2.

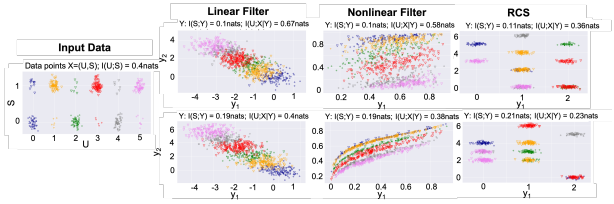

Figure 12. Results on synthetic data with $I(U;S) = 0.4$. From left to right: Input distribution (noise added for visual clarity); Linear filter results with tolerances $k = [0.9, 0.27]$; Nonlinear results at same tolerances; Representation of output distribution learned from the RCS sequences derived in Lemma 2.1.

Facial Images

Here we describe the U-Net-based architecture used to implement the obfuscation filter on all real data experiments. Figure 14 shows the network diagram, the presented architecture is fully convolutional, and the same network definition was used across all three experiments.

The filter network was first initialized as a passthrough network (trained to copy input image under RMSE loss). In the *subject-within-subject*, an additional output of the filter was trained to infer a class label on whether the subject was a consenting user or not. We stress that this was only done on the initialization phase of the *subject-within-subject* example, and that this class label was not explicitly preserved or retrained during the normal execution of Algorithm 3. The posterior inference networks are instances of Xception networks, shown in Figure 15.

All examples on real data were trained using Algorithm 3, where we impose the additional constraints of a fixed utility inference algorithm that the obfuscation filter must conform to; the sensitive attribute inference network was always trained adversarially. Training on a single Tesla K80 for 4 different tolerance values k under these conditions takes 3 to 6 days

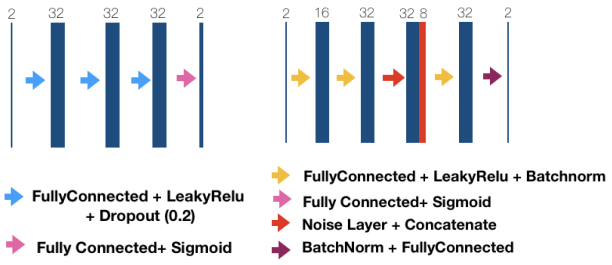

Figure 13. Left figure shows architecture of the utility and secret inference network used in the synthetic data examples. Right figure shows architecture of the nonlinear filter network used in the synthetic data examples

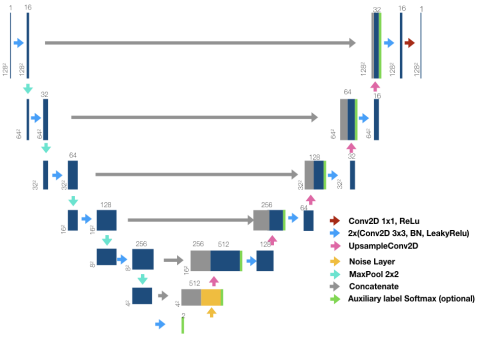

Figure 14. Obfuscation filter architecture based on U-Net ((Ronneberger et al., 2015)). There is a single noise layer (shown in yellow) where standard Gaussian noise is injected into the network to add stochasticity to the filter. The other notable component is the auxiliary label softmax, used for the *subject-within-subject* experiment. This extra layer was trained only during network initialization, but was not preserved during the final training stage. Input image sizes are shown for the *subject-within-subject* experiment.

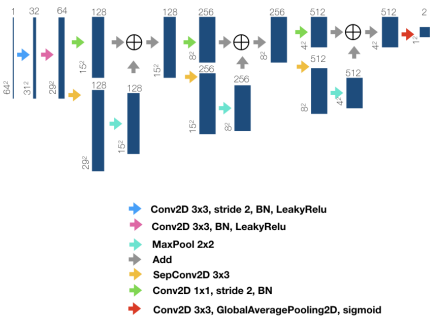

Figure 15. Architecture of utility and sensible variable inference networks used across all face experiments. These architectures are small modifications of Xception networks proposed in (Chollet, 2017)

Performance of Additional Attributes on Emotion vs Gender Images

We examine how other facial attributes such as face shape (Oval/Square) and age (Young/Old) are affected by the sanitization mapping where we wish to preserve gender identification and remove emotion identification. Results are shown in Table 4 for classifiers that are trained on normal images, and classifiers that are trained on the resulting sanitized images. Both attributes are neither spatially co-located or correlated with the obfuscated emotion attribute, this could potentially explain why these attributes are mostly unaffected by the learned data sanitization.

Table 4. Results across several tolerance parameters k. Confidence and accuracy results are shown for fixed and retrained classifiers for facial shape (Oval/Square) and age (Young/Old). The sanitization mapping was trained to obfuscate emotion information (Smiling/Non-smiling) while preserving gender information(Male/Female).

TOL K	FIXED SHAPE		RETRAINED SHAPE		FIXED AGE		RETRAINED AGE	
	CONF	ACC	CONF	ACC	CONF	ACC	CONF	ACC
∞	0.62	71.7%	0.62	71.7%	0.75	83.4%	0.75	83.4%
0.5	0.61	70.9%	0.61	71.0%	0.73	80.0%	0.74	82.4%
0.4	0.61	70.8%	0.60	71.3%	0.73	79.8%	0.73	82.0%
0.3	0.61	70.7%	0.61	71.1%	0.72	79.6%	0.73	81.3%
0.2	0.60	70.7%	0.60	71.0%	0.69	78.3%	0.71	80.3%
GUESS	0.52	60.7%	–	–	0.50	51.9%	–	–