Understanding Compressive Adversarial Privacy

Xiao Chen, Peter Kairouz, Ram Rajagopal

***Abstract*— Designing a data sharing mechanism without sacrificing too much privacy can be considered as a game between data holders and malicious attackers. This paper describes a compressive adversarial privacy framework that captures the trade-off between the data privacy and utility. We characterize the optimal data releasing mechanism through convex optimization when assuming that both the data holder and attacker can only modify the data using linear transformations. We then build a more realistic data releasing mechanism that can rely on a nonlinear compression model while the attacker uses a neural network. We demonstrate in a series of empirical applications that this framework, consisting of compressive adversarial privacy, can preserve sensitive information.**

I. Introduction

Machine learning has progressed dramatically in many real-life tasks such as classifying image [1], processing natural language [2], predicting electricity consumption [3], and many more. These tasks rely on large datasets that are usually saturated with private information. Data holders who want to apply machine learning techniques may not be cautious about what additional information the model can capture from training data, as long as the primary task can be solved by some model with high accuracy.

In this paper, we propose a privatization mechanism to avoid the potential exposure of the sensitive information while still preserving the necessary utility of the data that is going to be released. This mechanism largely leverages the concept of the game-theoretic approach by perturbing the data and retraining the model iteratively between data holder and malicious data attacker. Such a data perturbation idea is highly correlated with feature transformation and selection of the raw data input that correlated with private labels.

Protecting privacy has been extensively explored in myriad literature. A popular procedure is to anonymize the identifiable personal information in datasets (e.g. removing name, social security number, etc.). Yet anonymization doesn't provide good immunity against correlation attacks. A previous study [4] was able to successfully deanonymize watch histories in the Netflix Prize, a public recommender system competition. Another study designed re-identification attacks on anonymized fMRI (functional magnetic resonance imaging) imaging datasets [5]. On the other hand, the Differential privacy (DP) [6] has a strong standard of privacy guarantee and is applicable to many problems beyond database release [7]. This DP mechanism has been introduced in data privacy analysis in control and networks [8]–[11]. In particular, [8] gave a thorough investigation on performing the centralized and distributed optimization under differential privacy constraints. In this line of research, [10] and [11] focused on the cases of dynamic data perturbations in control systems. [9] presented a noise adding mechanism to protect the differential privacy of network topology.

However, training machine learning models with DP guarantees using randomized data often leads to a significantly reduced utility and comes with a tremendous hit in sample complexity [12], [13]. A recent work [14] applied the DP concept on a deep neural network to demonstrate that a modest accuracy loss can be obtained at certain worst-case privacy levels. However, this was still a "context-free" approach that didn't leverage the full structure between the data input and output.

To overcome the aforementioned challenges, we take a new holistic approach towards enabling private data publishing with consideration on both privacy and utility. Instead of adopting worst-case, context-free notions of data privacy (such as differential privacy), we introduce a context-aware model of privacy that allows the data holder to cleverly alter the data where it matters.

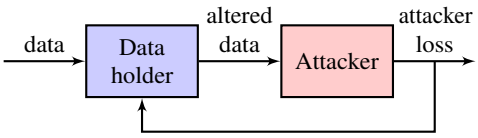

Fig. 1. Data releasing schematic

Our main contributions are listed as follows. First, with the goal of having a "distribution free" data releasing mechanism and inspired by general min-max games, we investigate a typical way of perturbing the data that is the compression using the data-driven approach. As a second contribution, we formulate the interaction between data holders and attackers through convex optimization during the min-max game when both players apply linear models. A corresponding equilibrium can be found and used as the optimal strategy for the data holder to yield the altered data. The third contribution is that our thorough evaluations of realistic datasets demonstrate the effectiveness of our compressive adversarial privacy framework. Finally, we leverage the mutual information to

X. Chen is with Department of Civil and Environmental Engineering, Stanford University, CA 94305, USA. markcx@stanford.edu

P. Kairouz is with Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043, Department of Civil and Environmental Engineering, and Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. kairouzp@stanford.edu

R. Rajagopal is with Department of Civil and Environmental Engineering, and Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. ramr@stanford.edu

This work is partially supported by NSF CAREER Award ECCS-1554178, NSF CPS Award #1545043 and DOE SunShot Office Solar Program Award.

validate that sensitive information can be protected from the privatized data.

The remainder of our paper is arranged as follows. In section II, we introduce the general adversarial privacy game. Section III describes the compressive adversarial privacy game with several cases of realistic data analyses. Section IV describes the quantification of privacy. Section V concludes the paper.

II. Privacy Preserved Releasing

We propose a general data publishing framework by incorporating the following game concept. In general there are two roles in this data releasing game: a data holder and a data consumer. Among data consumers, some people are good users who explore the pattern and extract the value of the data. We merge the good data consumers together with data holders and address their roles from the data holder perspective, since good data user are irrelevant in this game. Yet there are people who also try to learn personal sensitive information from the data on purpose. We define those malicious users as attackers. We focus on the data holders and attackers for the following description of the game.

Consider a dataset $\mathcal{D}$ which contains both the original data X and the associated customers' sensitive demographic information Y (e.g. account holder age, house square-footage, gender, etc.). Thus a sample i has a record $(x_i, y_i) \in \mathcal{D}$. We denote the function g as a general mechanism for a data holder to release the data. The released data are denoted as $\tilde{x_i}$ for the customer i, which can also be described as $\tilde{x_i} = g(x_i, y_i)$. Notice we don't release y_i to the public because it's private information. Generally speaking, $\tilde{X} = g(X, Y)$. Let the function h represent the adversarial hypothesis, e.g. the estimated outcome $\hat{Y} = h(\tilde{X})$. The attacker would like to minimize the inference loss on private labels, namely $\ell(\hat{Y}, Y)$ given some loss function ℓ, while the data holder would like to maximize the attacker's loss, and in principle, also wants to preserve the quality of the released data for research purposes. This data quality is characterized by some distance function measuring between the original data and the altered data. Therefore, we formulate a min-max game between the data holder and the attacker as follows:

$$\max_{g \in \mathcal{G}} \left\{ \min_{h \in \mathcal{H}} \ell\Big(h\big(g(X, Y)\big), Y\Big) \right\} \tag{1}$$

$$s.t. \quad d\big(X, g(X, Y)\big) \leq \gamma, \tag{2}$$

where $d()$ could be some distance function, such as Total Variation (TV), Wasserstein-1, or Frobenius norm, etc. [15], [16], and γ is a hyper-parameter. The constraint ensures that the released data will not be distorted too much from the original data.

This framework allows an attacker to incorporate any loss functions and design various adversarial inference models, which typically take the released data to predict the personal information. Given such a challenge, the data publisher has to design a good privatization mechanism g and γ to deteriorate the attacker's performance, which are also data dependent. For simplicity, we focus on the supervised learning setting in this work, but the concept can potentially be extended to the unsupervised learning.

III. Compressive Adversarial Privacy

A typical method to enforce data privacy is data compression. This method is well studied in [17] from a theoretical point of differential privacy. In reality, data compression is used in many applications such as text messaging and video transmission to protect the privacy. In this section, we extend the general min-max framework to a compression approach, namely a compressive adversarial privacy framework, as shown in Figure 2

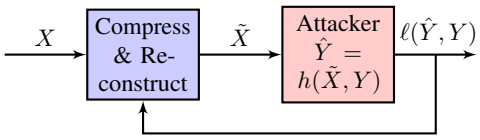

Fig. 2. Data compression schematic. $\tilde{X}$, which has the same dimension as X, is the reconstructed data that will be released to the public. The attacker infers the private labels Y by choosing good predictor h, getting the resulting $\hat{Y}$, and minimizing the inference loss denoted as $\ell(\hat{Y}, Y)$.

We focus on two scenarios to illustrate concrete privatization mechanisms. The first one is the linear compression when an attacker takes a linear model. The second one is the non-linear compression when an attacker uses a neural network. We evaluate of both cases based on real data.

A. Linear compression with continuous label

We introduce the case of an attacker who uses a linear model and the least squared loss function to infer private information through released data. This case is practical especially when private labels are continuous and have a linear relationship with the original data. We denote the original data matrix X and altered data matirx $\tilde{X} \in \mathbb{R}^{n \times p}$, where n is number of samples, p is number of features, and $\mathbb{R}$ is the set of real numbers. Such a data matrix contains individual samples x_i and $\tilde{x}_i \in \mathbb{R}^p$ respectively, where $i = 1, \ldots, n$. The private-info matrix $Y \in \mathbb{R}^{n \times d}$ consists of $y_i \in \mathbb{R}^d$ that each sample i has d types of private labels.

Consider a data holder who has a simple data-releasing mechanism that applies a linear transformation of X, namely projecting it down to lower dimensions to protect confidential information, i.e. $Z = XA$, where the matrix $A \in \mathbb{R}^{p \times k}, k < p$. Hence, $Z \in \mathbb{R}^{n \times k}$. In order to release meaningful data that still can be utilized by a majority of good users, the data holder performs a linear operation by multiplying $B \in \mathbb{R}^{k \times p}$ on Z and recovers it back to the same dimension as X, yielding $\tilde{X} = XAB$. The attacker fits a linear model to minimize the mean squared loss that is $\frac{1}{n}\sum_{i=1}^{n} \|\tilde{\Theta}^T \tilde{x}_i - y_i\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} \|\tilde{\Theta}^T B^T A^T x_i - y_i\|_2^2$, where $\tilde{\Theta} \in \mathbb{R}^{p \times d}$. Because the domain of $\tilde{\Theta}^T B^T$ is contained in Θ^T which is in $\mathbb{R}^{d \times k}$. This attacker's loss is lower bounded by

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \|\Theta^T A^T x_i - y_i\|_2^2, \tag{3}$$

where $\Theta \in \mathbb{R}^{k\times d}$. Therefore, when the data holder maximizes the attacker's loss, we can maximize this lower bound that automatically maximizes the minimum loss of the attacker. The resulting min-max problem can be formulated as

$$\max_{A,B} \min_{\Theta} \frac{1}{n}\sum_{i=1}^{n} \|\Theta^T A^T x_i - y_i\|_2^2 \tag{4}$$

$$s.t. \quad \|XAB - X\|_F^2 \le \gamma, \tag{5}$$

where $\|\cdot\|_F$ is the Frobenius norm. Given A, we further simplify the expression by finding the best recovering matrix $\hat{B}$ in place of B as follows (see VI-B for details):

$$\hat{B} = (A^T X^T X A)^{-1} A^T X^T X = (A^T A)^{-1} A^T = A^\dagger. \tag{6}$$

We denote the $A^\dagger$ to be the pseudo-inverse of A. The best predictor Θ for the attacker can be expressed as $\Theta = (A^T C_{xx} A)^{-1} A^T C_{xy}$, where $C_{xx} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T, C_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i^T$. Substituting Θ and B, we have the following problem:

$$\max_{A} \left\{ -\operatorname{Tr}\left(C_{xy}^T A (A^T C_{xx} A)^{-1} A^T C_{xy}\right)\right\} \tag{7}$$

$$s.t. \quad \|XA(A^T X^T X A)^{-1} A^T X^T X - X\|_F^2 \le \gamma. \tag{8}$$

Notice that $C_{xx} = \frac{1}{n} X^T X$. By flipping the sign of the maximization and denoting $M = A(A^T X^T X A)^{-1} A^T$ which is a positive semidefinite matrix (see the appendix VI-C), we have the following problem:

$$\min_{M} \frac{1}{n} \operatorname{Tr}\left(C_{xy}^T X M X^T C_{xy}\right) \tag{9}$$

$$s.t. \quad M \succeq 0 \tag{10}$$

$$\|XMX^T X - X\|_F^2 \le \gamma \tag{11}$$

$$rank(M) = k. \tag{12}$$

We put a rank constraint (12) because the dimension A is $p \times k, (k < p)$. This problem can be further relaxed to a convex optimization by regularizing the nuclear norm of matrix M as follows:

$$\min_{M} \frac{1}{n} \operatorname{Tr}\left(C_{xy}^T X M X^T C_{xy}\right) + \beta\|M\|_* \tag{13}$$

$$s.t. \quad \|XMX^T X - X\|_F^2 \le \gamma \tag{14}$$

$$M \succeq 0, \tag{15}$$

where $\|\cdot\|_*$ is the nuclear norm of a matrix that heuristically controls the rank of a matrix. Such a convex relaxation allows the data publisher to find an optimal solution of M, and correspondingly yields the appropriate $\tilde{X}$ (see appendix VI-A). Thus, both players can achieve an equilibrium in this game. To ensure the problem is feasible, one caveat is that we cannot pick arbitrarily small γ without considering the aforementioned rank k. We note that $\tilde{X} = XAA^\dagger$ is a low rank-k approximation of the original data matrix X.

Theorem 1: Suppose a rank-p matrix X consists of the singular values $\lambda_1, \ldots, \lambda_p$. With the best rank-$k$ approximation $\tilde{X}_k$ under Frobenius norm, the distortion threshold γ is at least $\sum_{i=k+1}^{p} \lambda_i^2$.

We put the proof in appendix VI-F. The theorem reveals the relationship between setting the distortion tolerance γ and the rank k. Hence, a simple algorithm (Algorithm 1) is proposed for the data holder to generate $\tilde{X}$.

Algorithm 1 Generating $\tilde{X}$ (Linear attacker)

Input: dataset $(X, Y) \in \mathcal{D}$, parameter γ, β_0, k, η.
Output: $\tilde{X}$
partition dataset into several batches (X, Y).
for a batch of $(X, Y) \in \mathcal{D}$ **do**
  $\hat{k} = 0$, $t = 0$
  while $\hat{k} \neq k$ **do**
    $M_t = \arg\min_M \{\frac{1}{n}\operatorname{Tr}(C_{xy}^T X M X^T C_{xy}) + \beta_t\|M\|_*$ s.t. $\|XMX^T X - X\|_F^2 \le \gamma\}$ (solving the optimization in equation (13 - 15) with certain values of γ, β_t, k).
    $U_t, \Lambda_t \leftarrow SVD(M_t)$ (applying Singular Value Decomposition on $M_t = U_t \Lambda_t U_t^T$ to get matrices U_t and Λ_t.)
    $\hat{k} = rank(\Lambda_t)$. check the rank of the matrix Λ with non trivial eigenvalues. (e.g $\lambda_j > \eta\lambda_{max}$, where $\forall j = 1, \ldots, n; \eta = 0.01$.)
    if $\hat{k} = k$ **then**
      break
    else if $\hat{k} > k$ **then**
      $\beta_{t+1} \leftarrow \beta_t + \frac{\beta_t}{2}$
    else if $\hat{k} < k$ **then**
      $\beta_{t+1} \leftarrow \beta_t - \frac{\beta_t}{4}$
    end if
    $t = t + 1$
  end while
  $\tilde{X} \leftarrow XMX^T X$
end for

Remark 1: This approach can be interpreted as releasing a low dimensional approximation to a set of data, incorporating the relation between the original data and private labels, while still maintaining a certain distortion between the released data and the original data.
We also discovered that a similar scheme can be applied on compressing original data with additive Gaussian noise. See Appendix VI-E for details.

B. Case study: Power consumption data

The first experiment of our analysis uses the CER dataset, which was collected during a smart metering trial conducted in Ireland by the Irish Commission for Energy Regulation (CER) [18]. The dataset contains measurements of electricity consumption gathered from over 4000 households every 30 minutes between July 2009 and December 2010. Each participating household was asked to fill out a questionnaire about the households' socio-economic status, appliances stock, properties of the dwelling, etc. [19]. To demonstrate our concepts, we sampled a portion of the customers who has valid entries of demographic information, e.g. number of

appliances and floor area of the individual house. In the following experiment, we treat floor area as private data Y.

Throughout the case simulations, we extract the four-week time series in September 2010. Since the power consumption (in kilowatts) is recorded every 30 minutes, there are $2 \times 24 \times 28 = 1344$ entries for a single household. To simplify the input dimension and avoid the over-fitting issue from raw input, we compute a set of features on the electricity consumption records of a household. The features then serve as the input to the prediction model. Table IV lists all 23 features we calculated from electricity consumption data, which is also used in [19]. We treat these features as X and normalize them such that they range from 0 to 1. Data normalization is required in our experiment in that it gets rid of the scale inconsistency across the different features.

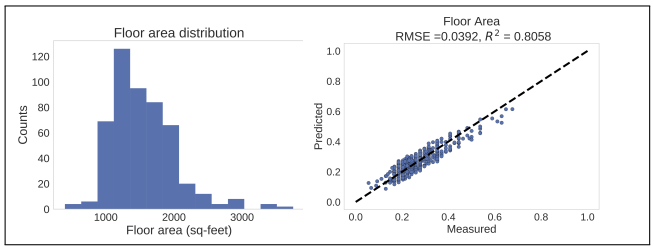

Fig. 3. Display the sampled data. **Left panel** shows the histogram of *Floor area* of each house. **Right panel** shows the prediction capability of a linear model that fits on the *Floor area*

In the linear transformation model, given Y is the private information, we run algorithm 1 to release $\tilde{X}$. This procedure involves solving semidefinite programming, which could be slow when the dimension of input samples is large. So we partition the samples into several groups with a reasonable number of households in each group (e.g. 30 to 40 as long as the number of households is larger than the number of features). After running experiments on several rank conditions of data matrices, we found that lower rank indicates better privacy (higher prediction error), given in Table I. With a low rank condition that the data holder maintains, the attacker can barely (see Table I) predict the private label Y. We also partitioned the data into 80% for training and 20% for testing. Table I shows the corresponding results with different ranks of the compression matrix for the testing set. A batch of released data differs from the original when rank is $4, 10$ and 18. The difference is shown in Figure 4.

TABLE I

METRICS OF LINEAR TRANSFORMED DATA

Rank	RMSE	R^2	distortion
4	6.7e+04	1.61e-08	0.616
10	7.1e+01	2.28e-03	0.081
18	8.7e-01	1.12e-02	0.079
23	3.9e-02	8.01e-01	0

C. Nonlinear compression with categorical variable

Another common type of data has publishable features X are high-dimensional continuous and the private labels Y are discrete, for instance, images with some discrete labels (e.g. gender). Generally speaking, a sample i has $y_i \in \mathcal{Y} = \{-1, +1\}$ and $x_i \in \mathbb{R}^p$ where x_i^T is the ith row of the data matrix X. The data holder designs a nonlinear compression mechanism to reduce the classification accuracy of y_i given $\tilde{x}_i$, where $\tilde{x}_i = g(x_i, y_i)$. We assume the attacker can use an advanced model, e.g. neural networks, to estimate the private labels. We further specify that h and g are functions parametrized by θ_h and θ_g. The attacker minimizes the estimation loss, that is, $\min_{\theta_h} \ell(h_{\theta_h}(g_{\theta_g}(X, Y)), Y)$. The data holder designs a compressive function g to maximize the attacker's loss as well as maintain a certain distortion γ as aforementioned in equations (1),(2). This min-max game is difficult to find its equilibrium point in the context of neural networks with constraints, because the objective functions are non-convex with respect to parameters. Therefore, we use a heuristic way to cast the constrained optimization into a unconstrained optimization with regularization as follows:

$$\max_{\theta_g} \Big\{ \min_{\theta_h} \frac{1}{n} \sum_{i=1}^{n} \ell\Big(h_{\theta_h}\big(g_{\theta_g}(x_i, y_i)\big), y_i\Big) \tag{16}$$

$$-\beta\big((\frac{1}{n}\sum_{i=1}^{n} \|g_{\theta_g}(x_i, y_i) - x_i\|^2) - \gamma\big)^2 \tag{17}$$

$$+\rho \min\{0, \gamma - (\frac{1}{n}\sum_{i=1}^{n} \|g_{\theta_g}(x_i, y_i) - x_i\|^2)\}\Big\}, \tag{18}$$

where β and ρ are the hyper parameters controlling the iterates satisfied by the constraints. The distortion is characterized by the averaged Euclidean norm of the difference in samples. We propose a simple min-max alternative algorithm (Algorithm 2) to obtain the parameter θ_g for the function g and yield the corresponding $\tilde{X}$. Similar to the idea of the Augmented Lagrangian method [20], the scale of β and ρ are gradually increasing as the iteration step increases. The term (18) is added to ensure the solution strictly satisfies the constraint mentioned in expression (2). Other alternative approaches are also proposed in [21]–[23]. Distinguished from those works, we construct a convex approximation with distortion constraints that is applied in privacy games.

D. Case study: Images of people

To perform our experiment of the nonlinear compressive model with a categorical response variable, we use the Groups of People dataset [24]. The dataset contains 4550 images from Flicker of human faces with labeled attributes such as age and gender. These images are 61×49 in grayscale pixels ranging from 0 to 255, with 3500 training and 1050 testing samples respectively. In this experiment, the images are X and the label of gender, which is evenly spread in both the training and testing sets, is Y. We label female or male as 1 or -1. Sampled raw images are shown in Appendix VI-H.

For the data holder to perform nonlinear compression, we implement a three-layer neural network, which shares the similar concept of the autoencoder [25]. The first two layers serve as an encoder. The initial layer has 2989 units that takes original vectorized images [$2989 = 61 \times 49$], followed by a ReLU activation and batch normalization. We vary the second layer units from $2048, 512$, and 128 for several cases, which

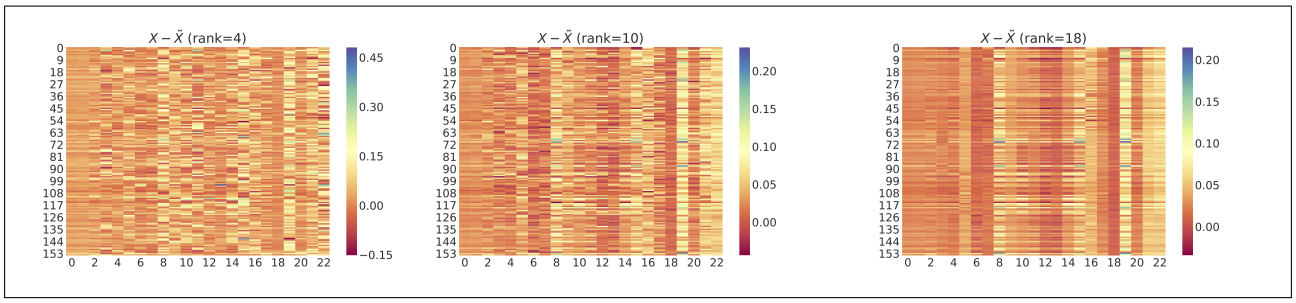

Fig. 4. Difference between altered and original data when rank equals 4, 10, and 18

Algorithm 2 Generating $\tilde{X}$ (Neural Net attacker)

1: *Input:* dataset $\mathcal{D}$, parameter γ, iteration number T
2: *Output:* Optimal data publisher parameters θ_g
3: Initialize θ_g^t and θ_h^t when $t = 0$
4: **for** $t = 0, ..., T$ **do**
5: take minibatch of n samples $\{x_{(1)}, \ldots, x_{(n)}\}$ drawn randomly from $\mathcal{D}$
6: Generate $\tilde{x}_{(i)} = g_{\theta_g}(x_{(i)}, y_{(i)})$ for $i = 1, \ldots, n$
7: Compute the parameter θ_h^{t+1} for the adversary $\theta_h^{t+1} = \arg\min_{\theta_h} \frac{1}{n} \sum_{i=1}^{n} \ell(h_{\theta_h}(\tilde{x}_{(i)}), y_{(i)})$
8: Compute the descent direction $\nabla_{\theta_g} \mathcal{L}(\theta_g, \theta_h^{t+1})$, where

$$\mathcal{L}(\theta_g, \theta_h^{t+1}) = -\frac{1}{n} \sum_{i=1}^{n} \ell(h_{\theta_h^{t+1}}(g_{\theta_g}(x_{(i)}, y_{(i)})), y_{(i)})$$

$$+\beta\big((\frac{1}{n} \sum_{i=1}^{n} \|g_{\theta_g}(x_{(i)}, y_{(i)}) - x_{(i)}\|_2^2) - \gamma\big)^2$$

$$+\rho \max\{0, (\frac{1}{n} \sum_{i=1}^{n} \|g_{\theta_g}(x_i, y_i) - x_i\|^2) - \gamma\}$$

9: Perform backtracking line search along $\nabla_{\theta_g} \mathcal{L}(\theta_g, \theta_h^{t+1})$ and update $\theta_g^{t+1} = \theta_g^t - \alpha_t \nabla_{\theta_g} \mathcal{L}(\theta_g, \theta_h^{t+1}), \quad \alpha_t > 0$
10: Exit if solution converged
11: **end for**
12: **return** θ_g^{t+1}; $\tilde{X} = g_{\theta_g^{t+1}}(X, Y)$

are denoted as compression-rank. We define the corresponding compression-rank rate 0.685, 0.171, and 0.043 to be *high*, *medium* and *low* respectively[1]. The last layer, connected with ReLU activation, has the same dimension as the vectorized image input that performs the role of a decoder. The attacker is represented by a 3-layer neural network, comprised of an initial 2989 units layer, followed by 2048 units layer, and lastly a two units layer as softmax output. We apply leaky ReLU activation and batch normalization between each layer.

Before considering adversarial compression, we first classify reconstructed images with different compression-rank rates without having a min-max game. This operation serves two purposes: a) investigating the accuracy of gender classification; b) fetching the minimum distortion threshold γ in the context of mean squared error loss (i.e. min $\frac{1}{n} \sum_{i=1}^{n} \|\tilde{x}_i - x_i\|^2$ yields the smallest γ). The following results are evaluated based on the testing set. Figure 5 displays a sampled image associated with different scenarios. A lower compression rank rate yields a worse image quality. Table II shows that compressing images with the high and medium ranks doesn't reduce the gender classification accuracy too much, yielding a relatively low image quality loss. In the example of high compression-rank, the average distortion per pixel is $0.0166 * 10^2 \approx 1.6\%$ which is not too large.

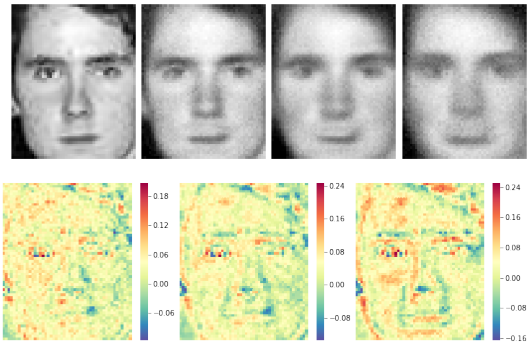

Fig. 5. A sampled face. **Top row**: From left to right are an original image, and then decoded images with *high, medium* and *low* compression-rank rate. **Bottom row**: From left panel to right panels are the result of the pixel difference between decoded and raw image projected to 0-255 with *high*, *medium* and *low* compression-rank rates respectively

TABLE II
CLASSIFICATION RESULTS OF GENDER WITH THE RAW DATA UNDER DIFFERENT COMPRESSION-RANK CASES

compression-rank	accuracy(gender)	distortion/pixel	distortion
raw (2989)	0.692	0	0
high (2048)	0.685	0.0166	0.195
medium (512)	0.664	0.0259	0.304
low (128)	0.627	0.0312	0.365

Utilizing the previous result as a reference, we pick several proper values of γ to further understand the adversarial privacy compression. In the high compression-rank case, we test three scenarios where γ is $0.3, 2$ and 4 respectively. We discover that the encoder-decoder tends to alternate pixels near eyes, mouths, and rims of hair. A similar patten can also be observed when we test the low compression-rank

[1]the compression-rank rate is obtained by number of bottleneck units divided by input units. e.g. $\frac{2048}{2989} = 0.685$

case where γ is $1, 2$ and 4. We also notice that the low compression-rank scenario has a more scattered dotted patten of black/gray pixels at the large tolerance level, whereas the high compression-rank case has more concentrated black pixels, as shown in Figure 6. We believe the reason is that the data holder always adjusts the pixels that are highly correlated with gender. Since the high compression-rank encoder-decoder preserves more information than the low compression-rank one, it's much easier for the data holder to alter the target pixel features within limited total distortion. The privatized images generated through min-max training indeed yield lower prediction accuracy of gender than the original encoded-decoded images. Table III depicts the gender classification results indicating that it is harder to predict gender with increased distortion. The table also reveals that higher compression rank performs better in terms of decreasing the accuracy if the distortion is sufficiently large.

TABLE III

CLASSIFICATION ACCURACY OF GENDER WITH THE DATA RELEASED UNDER ADVERSARIAL PRIVACY

compression-rank	$\gamma = 0.3$	$\gamma = 1$	$\gamma = 2$	$\gamma = 4$
high (2048)	0.628	0.600	0.573	0.486
medium (512)[1]		0.607	0.594	0.512
low (128)		0.602	0.585	0.521

[1] $\gamma = 0.3$ is unattainable, since the compression rank is small enough so that the minimum reconstruction loss (Mean Squared Error) already reachs to the 0.3.

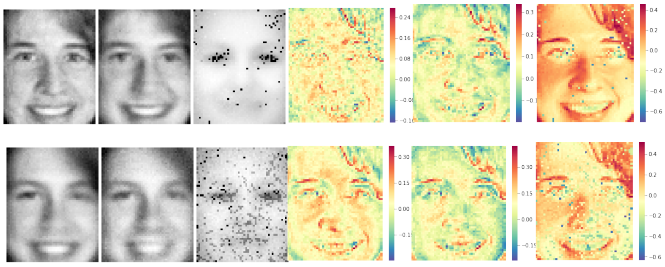

Fig. 6. A sampled image with different compression rank and distortion tolerance γ. **Top left**: From left to right are visual output when γ equals 0.3, 2, 4 in the high compression-rank case. **Top right**: From left to right, these images show the difference between output images and raw images for corresponding γ value in the high compression-rank case. **Bottom left**: From left to right are visual output when γ equals 1, 2, 4 in the low compression-rank case. **Bottom right**: From left to right are difference between output images and raw images for corresponding γ value in the low compression-rank case.

IV. PRIVACY GUARANTEE

Our previous experiments show that a (local) equilibrium can be achieved through this min-max game approach. While we cannot preclude that there may be some other equilibria in the context of a neural network. Thus, a quantifiable metric is needed to give privacy guarantee between sensitive data and altered data.

In this section we introduce the empirical mutual-information concept to quantify the privatization quality of this min-max game approach, i.e. measuring the correlation between the sensitive response data and the released feature data pre- and post-privatization. Mutual information (MI) [26] is a well established tool that has been widely adopted to quantify the correlation between the two streams of data by a non-negative scalar [27]. From the data driven perspective, we have the empirical MI $\hat{I}(X;Y) = \hat{H}(X) - \hat{H}(X|Y)$, where $\hat{H}$ characterizes the empirical entropy. This empirical entropy can be calculated using the classical nearest k-th neighbor method [28].

Continuous response label: Given that Y is continuous, the mutual information can be expressed as $\hat{I}(X;Y) = \hat{H}(X) + \hat{H}(Y) - \hat{H}(X,Y)$, where $\hat{H}(X), \hat{H}(Y)$ can be obtained directly from method in [28], given samples x_i, y_i. The joint empirical entropy is calculated by concatenating each x_i and y_i together as one sample and using the nearest neighbor entropy estimation again.

Categorical response label: For the discrete response $Y \in \{-1, +1\}$, we have the mutual information $\hat{I}(X;Y) = \hat{H}(X) - \hat{H}(X|Y) = \hat{H}(X) - \big(p(Y = -1)\hat{H}(X|Y = -1) + p(Y = 1)\hat{H}(X|Y = 1)\big)$, where $p(Y = \pm 1)$ can be approximated by the sample frequency in the dataset, and $\hat{H}(X|Y = \pm 1)$ can be obtained by the aforementioned k-th nearest neighbor method with partitioned samples according to the value of Y.

For the experiment of the continuous response variable, we first calculate the empirical MI between the power consumption statistics and floor areas. The original MI between power usage statistics data and floor area is $\hat{I}(X;Y) = 2.150$. The resulting MI between altered power usage data and floor areas, which is denoted by $\hat{I}(\tilde{X};Y)$, are 0.995, 0.494, and 0.216 when the rank of compression matrices are 18, 10, and 4. For the categorical response variable experiment, the empirical MI between the images data and gender data is obtained as follows. The original MI between raw images X and gender label Y is $\hat{I}(X;Y) = 0.249$. When we pick high compression-rank with $\gamma = 0.3, 1, 2$ and 4, the $\hat{I}(\tilde{X};Y)$ are $0.217, 0.170, 0.105$, and 0.012. The medium compression-rank yields $\hat{I}(\tilde{X};Y)$ to be $0.179, 0.112$, and 0.014 for $\gamma = 1, 2$, and 4 respectively. In the low compression rank case, $\hat{I}(\tilde{X};Y)$ are $0.174, 0.101$, and 0.017 with the aforementioned γ. We notice the empirical MI indeed decreases as the distortion increases. Due to the challenge of high dimensional data, we apply the principal component analysis to project the $\tilde{X}$ down to 16 dimensions and find the approximate $\hat{I}(\tilde{X};Y)$. This is an alternative attempt to demonstrate the effectiveness of using our framework. The changes of mutual information value show the privacy guarantee between the released data and sensitive labels under various distortion conditions. Yet we believe a more advanced architecture of the neural network can be applied to extract the embeddings of semantic features, resulting a better estimates of empirical mutual information. We will explore this potential direction in our future research.

V. CONCLUSION

Recent breakthroughs in artificial intelligence require a huge amount of data to support the learning quality of various models. Yet the risk of data privacy is often overlooked in the

current data sharing processes. The recent news of data leakage by Facebook shows that privacy risk could significantly impact some issues in politics. Therefore, securely designing a good data privatization mechanism is important in the context of utilizing machine learning models. By thorough evaluations, our new min-max adversarial compressive privacy framework provide an effective and robust approach to protect private information. We leverage the data-driven approach without posing assumptions on data distribution. It's crucial during practical implementation, since the real data is often more complicated than a simple characterization of a parametric probability distribution. Along this line of research, many interesting extensions can be built on our framework to create a robust privacy protector for data holders.

VI. Appendix

A. Casting linear problem

Consider the following problem

$$\min_B \|XAB - X\|_F^2, \tag{19}$$

where matrix X is $n \times p$, matrix A is $p \times k$, and matrix B is $k \times p$. We give a brief minimizer derivation as follows:

$$\begin{aligned}\|XAB - X\|_F^2 &= \mathbf{Tr}\Big((XAB - X)(XAB - X)^T\Big) \\ &= \mathbf{Tr}\Big(XABB^TA^TX^T - 2XABX^T + XX^T\Big).\end{aligned}$$

The derivative of the first term with respect to B is

$$\begin{aligned}&\frac{\partial}{\partial B}\mathbf{Tr}(XABB^TA^TX^T) = \frac{\partial}{\partial B}\mathbf{Tr}(BB^TA^TX^TXA) \\ &= (A^TX^TXA)B + (A^TX^TXA)^TB = 2(A^TX^TXA)B.\end{aligned}$$

The derivative of the second term with respect to B is

$$\frac{\partial}{\partial B}\mathbf{Tr}(2XABX^T) = 2\frac{\partial}{\partial B}\mathbf{Tr}(BX^TXA) = 2(X^TXA)^T$$

Thus, we set the derivative equals zero and obtain the minimizer B as follows:

$$\begin{aligned}&\frac{\partial}{\partial B}\mathbf{Tr}\Big(XABB^TA^TX^T - 2XABX^T + XX^T\Big) && (20)\\ &= 2(A^TX^TXA)B - 2(X^TXA)^T \triangleq 0 && (21)\\ &\implies (A^TX^TXA)B = A^TX^TX && (22)\\ &\implies B = (A^TX^TXA)^{-1}A^TX^TX. && (23)\end{aligned}$$

Now we design the $\tilde{X}$ such that

$$\begin{aligned}\tilde{X} = XAB &= XA\underbrace{(A^TX^TXA)^{-1}A^TX^TX}_{B} && (24)\\ &= X\underbrace{A(A^TX^TXA)^{-1}A^T}_{M}X^TX. && (25)\end{aligned}$$

Instead of explicitly designing a low rank matrix A , we solve an alternative equivalent problem of determining a low rank matrix M to compress the data.

B. Recovering Linear Operation

Claim: B is pseudoinverse of A, i.e. $B = (A^TA)^{-1}A^T = A^\dagger$.

Proof: We apply Singular Value Decomposition (aka SVD, which is similar to PCA) on data matrix $X = USV^T$. U is $n \times k$, S is $k \times k$, V is $p \times k$. We have

$$\begin{aligned}B &= (A^TX^TXA)^{-1}A^TX^TX \\ &= \big(A^T(USV^T)^T(USV^T)A\big)^{-1}A^T(USV^T)^T(USV^T) \\ &= (A^TVSU^TUSV^TA)^{-1}A^TVSU^TUSV^T \\ &= (A^TVS\underbrace{U^TU}_{I}SV^TA)^{-1}A^TVS\underbrace{U^TU}_{I}SV^T \\ &= (A^TVS^2V^TA)^{-1}A^TVS^2V^T \\ &= (\underbrace{S^2A^TV}_{\text{swapped}}V^TA)^{-1}\underbrace{S^2A^TV}_{\text{swapped}}V^T \\ &= (S^2A^T\underbrace{VV^T}_{I}A)^{-1}S^2A^T\underbrace{VV^T}_{I} \\ &= (S^2A^TA)^{-1}S^2A^T = (A^TA)^{-1}A^T.\end{aligned}$$

C. Proof of positive semidefinite property of a matrix

Claim: $M = A(A^TX^TXA)^{-1}A^T$ is Positive Semidefinite. *Proof:* show A^TX^TXA is positive semidefinite. Since $A \in \mathbb{R}^{p\times k}, X \in \mathbb{R}^{n\times p}$, for any vector $v \in \mathbb{R}^k$, we have

$$v^TA^TX^TXAv = (XAv)^T(XAv) = \|XAv\|_2^2 \geq 0.$$

Therefore, we can apply Signaler Value Decomposition on (A^TX^TXA), we get $(A^TX^TXA) = VSV^T$, where $S = \mathrm{diag}(\sigma_1, \sigma_2, ..., \sigma_k)$. The resulting $(A^TX^TXA)^{-1}$ can be expressed as $(A^TX^TXA)^{-1} = V(S^{-1})V^T$ where $S^{-1} = \mathrm{diag}(1/\sigma_1, 1/\sigma_2, ..., 1/\sigma_k)$. Because all σ are positive, we denote $\delta_i^2 = \frac{1}{\sigma_i}$. Hence $\Delta = \mathrm{diag}(\delta_1, ..., \delta_k)$. And

$$M = AV\Delta\Delta^TV^TA^T. \tag{26}$$

For any $v \in \mathbb{R}^p$, we have

$$\begin{aligned}v^TMv &= v^T(\Delta^TV^TA^T)^T\Delta^TV^TA^Tv && (27)\\ &= \|\Delta^TV^TA^Tv\|_2^2 \geq 0. && (28)\end{aligned}$$

Thus M is positive semidefinite.

D. Convexity of a re-parameterized problem

Claim: The following optimization is convex:

$$\begin{aligned}&\min_M \mathrm{Tr}\left(C_{xy}^TXMX^TC_{xy}\right) + \beta\|M\|_* && (29)\\ &\text{s.t.} \quad \|XMX^TX - X\|_F^2 \leq \gamma. && (30)\end{aligned}$$

Proof: It is easy to see that the first term $\mathrm{Tr}\left(C_{xy}^TXMX^TC_{xy}\right)$ is convex, since M is positive semidefinite, trace operator is linear with respect to M. The second term and third term are also convex. For any norm, given $0 \leq \alpha \leq 1$ and two matrices A, B, we have

$\|\alpha A + (1-\alpha)B\| \leq \|\alpha A\| + \|(1-\alpha)B\| = \alpha\|A\| + (1-\alpha)\|B\|$

Hence Frobenius norm and Nuclear norm are specific forms of norm that is convex with respect to M. The first term in the objective is just linear in M. Thus, the problem is convex.

E. Deriving linear compression with noise

Consider the case $\varepsilon \sim \mathcal{N}(0, \Sigma)$. We have the min-max game as follows

$$\max_{A,B,\Sigma} \min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\varepsilon \sim P_\varepsilon} \|\Theta^T (A^T x_i + \varepsilon) - y_i\|_2^2 \tag{31}$$

$$- \gamma \|XAB - X\|_F^2 \tag{32}$$

For attacker, we have the following minimization problem

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\varepsilon \sim P_\varepsilon} \|\Theta^T (A^T x_i + \varepsilon) - y_i\|_2^2 \tag{33}$$

$$= \min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \Big((A^T x_i + \varepsilon_i)^T \Theta \Theta^T (A^T x_i + \varepsilon_i) \tag{34}$$

$$- 2(A^T x_i + \varepsilon_i)^T \Theta y_i + y_i y_i^T \Big) \tag{35}$$

We first find the minimizer Θ for the attacker. By taking the derivative over Θ, we have

$$\frac{1}{n} \sum_{i=1}^{n} \Big(A^T x_i x_i^T A + \varepsilon_i \varepsilon_i^T \Big) \Theta = A^T \frac{1}{n} \sum_{i=1}^{n} x_i y_i^T \tag{36}$$

$$\Theta = (A^T X^T X A + \Sigma)^{-1} A^T C_{xy} \tag{37}$$

Also we also find the best recover matrix B by considering the following relation $\tilde{X} = XA + \boldsymbol{\varepsilon}$

$$E_{\varepsilon \sim P_\varepsilon} \|\tilde{X} - X\|_F^2 \tag{38}$$

$$= \mathrm{Tr} \Big((XAB + \boldsymbol{\varepsilon} B - X)(XAB + \boldsymbol{\varepsilon} B - X)^T \Big) \tag{39}$$

$$= \mathrm{Tr}(XABB^T A^T X^T + B\Sigma B^T - 2XABX^T) \tag{40}$$

Taking the derivative over B and set it equals 0, we have $B = (A^T X^T X A + \Sigma)^{-1} A^T X^T X$. Hence the data holder's maximization can be casted into

$$\min_{A,\Sigma} \mathrm{Tr} \Big(C_{xy}^T A (A^T X^T X A + \Sigma)^{-1} A^T C_{xy} \Big) \tag{41}$$

$$+ \|XA(A^T X^T X A + \Sigma)^{-1} A^T X^T X - X\|_F^2 \tag{42}$$

It is not difficult to discover that $A(A^T X^T X A + \Sigma)^{-1} A^T$ is also positive semidefinite. Thus the problem can be relaxed to convex optimization.

F. Proof of Theorem 1

Proof: Denote $X = \sum_{i=1}^{p} \lambda_i u_i v_i^T$, where λ_i is singular value, u_i, v_i are corresponding left and right singular vectors. The best rank-k approximation $\tilde{X}_k = \sum_{i=1}^{k} \lambda_i u_i v_i^T$ is achieved by SVD in Frobenius norm by Eckart-Young theorem [29]. Then $\|X - \tilde{X}_k\|_F^2 = \mathrm{Tr}\left((\sum_{i=k+1}^{p} \lambda_i u_i v_i^T)(\sum_{i=k+1}^{p} \lambda_i u_i v_i^T)^T \right) = \mathrm{Tr}(\sum_{i=k+1}^{p} \lambda_i^2) = \sum_{i=k+1}^{p} \lambda_i^2$

G. Features extracted from power data

Features for power consumptions are displayed in Table IV.

H. Images

Original people images are shown in Figure 7

TABLE IV

FEATURES EXTRACTED OUT OF POWER CONSUMPTIONS

Index	Description
1	Week total mean
2	Weekday total mean
3	Weekend total mean
4	Day (6am - 10pm) total mean
5	Evening (6pm - 10pm) total mean
6	Morning (6am - 10am) total mean
7	Noon (10am - 2pm) total mean
8	Night (1am - 5am) total mean
9	Week max power
10	Week min power
11	ratio of Mean over Max
12	ratio of Min over Mean
13	ratio of Morning over Noon
14	ratio of Noon over Day
15	ratio of Night over Day
16	ratio of Weekday over Weekend
17	proportion of time with $P_t > 0.5kw$
18	proportion of time with $P_t > 1kw$
19	proportion of time with $P_t > 2kw$
20	sample variance of P_t
21	sum of difference $\lvert P_t - P_{t-1} \rvert$
22	sample cross correlation of subsequent days
23	number of counts that $\lvert P_t - P_{t-1} \rvert > 0.2kw$

Fig. 7. The sampled images with the dimension of 61×49 for each one.

I. Low rank linear transformation

Singular value matrices for a batch of samples with the low rank transformation are shown in Figure 8. And the raw and altered power consumption features are shown in Figure 9

REFERENCES

[1] B. Oshri, A. Hu, P. Adelson, X. Chen, P. Dupas, J. Weinstein, M. Burke, D. Lobell, and S. Ermon, "Infrastructure quality assessment in africa using satellite imagery and deep learning," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery; Data Mining*, ser. KDD '18. New York, NY, USA: ACM, 2018.

[2] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[3] R. Sevlian and R. Rajagopal, "A scaling law for short term load forecasting on varying levels of aggregation," *International Journal of Electrical Power & Energy Systems*, vol. 98, pp. 350–361, 2018.

[4] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008, pp. 111–125.

[5] E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable, "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity," *Nature neuroscience*, vol. 18, no. 11, pp. 1664–1671, 2015.

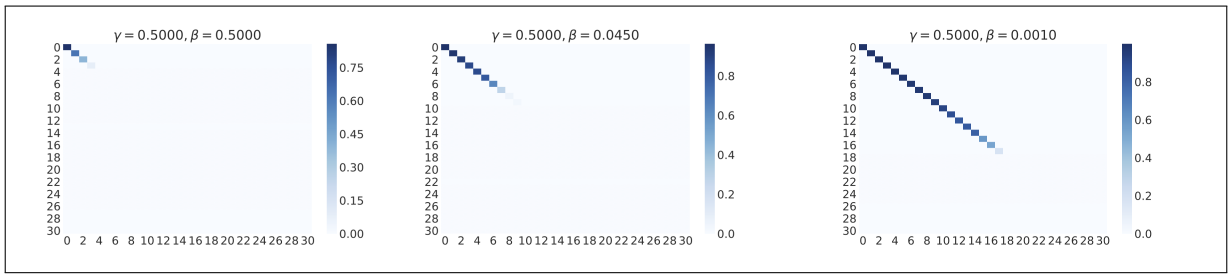

Fig. 8. Singular values for a sampled data patch when the rank equals 4, 10, and 18

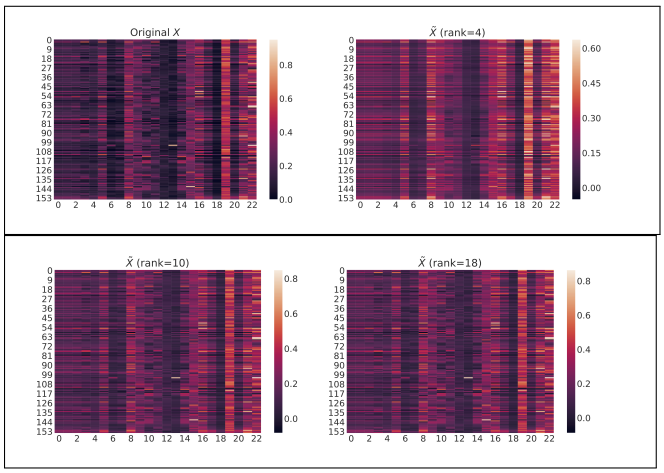

Fig. 9. The data holder applies a linear transformation on the data, i.e. compresses it down to low rank matrices and recovers it back to the original dimension. Each figure consists of 155 sampled households with 23 features which has been normalized originally. We fix the distortion tolerance γ and tune the nuclear norm coefficient β to get the result of various rank scenarios.

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

[7] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[8] J. Cortés, G. E. Dullerud, S. Han, J. Le Ny, S. Mitra, and G. J. Pappas, "Differential privacy in control and network systems," in *Decision and Control (CDC), 2016 IEEE 55th Conference on*. IEEE, 2016, pp. 4252–4272.

[9] V. Katewa, A. Chakrabortty, and V. Gupta, "Protecting privacy of topology in consensus networks," in *American Control Conference (ACC), 2015*. IEEE, 2015, pp. 2476–2481.

[10] Z. Huang, Y. Wang, S. Mitra, and G. Dullerud, "Controller synthesis for linear time-varying systems with adversaries," *arXiv preprint arXiv:1501.04925*, 2015.

[11] F. Koufogiannis and G. J. Pappas, "Differential privacy for dynamical sensitive data," in *Decision and Control (CDC), 2017 IEEE 56th Annual Conference on*. IEEE, 2017, pp. 1118–1125.

[12] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 2013, pp. 429–438.

[13] S. E. Fienberg, A. Rinaldo, and X. Yang, "Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables," in *International Conference on Privacy in Statistical Databases*. Springer, 2010, pp. 187–199.

[14] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 308–318.

[15] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "On surrogate loss functions and f-divergences," *The Annals of Statistics*, pp. 876–904, 2009.

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

[17] S. Zhou, K. Ligett, and L. Wasserman, "Differential privacy with compression," in *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*. IEEE, 2009, pp. 2718–2722.

[18] C. for Energy Regulation (CER), "Cer smart metering project - electricity customer behaviour trial, 2009-2010 [dataset]," *Irish Social Science Data Archive. SN: 0012-00. 1st Edition.*, 2012. [Online]. Available: www.ucd.ie/issda/CER-electricity

[19] C. Beckel, L. Sadamori, T. Staake, and S. Santini, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, pp. 397–410, 2014.

[20] C. Wu and X.-C. Tai, "Augmented lagrangian method, dual methods, and split bregman iteration for rof, vectorial tv, and high order models," *SIAM Journal on Imaging Sciences*, vol. 3, no. 3, pp. 300–339, 2010.

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

[22] J. Hamm, "Minimax filter: Learning to preserve privacy from inference attacks," *arXiv preprint arXiv:1610.03577*, 2016.

[23] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Context-aware generative adversarial privacy," *Entropy*, vol. 19, no. 12, 2017.

[24] A. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. CVPR*, 2009.

[25] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[26] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[27] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[28] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[29] G. H. Golub, A. Hoffman, and G. W. Stewart, "A generalization of the eckart-young-mirsky matrix approximation theorem," *Linear Algebra and its applications*, vol. 88, pp. 317–327, 1987.