An Improved Scheme for Privacy-Preserving Collaborative Anomaly Detection

Lingjuan Lyu[1] Yee Wei Law[2] Sarah M. Erfani[3] Christopher Leckie[3] Marimuthu Palaniswami[1]
[1] Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, Australia
[2] School of Engineering, University of South Australia, Australia
[3] Department of Computing and Information Systems, The University of Melbourne, Australia
Email: llv@student.unimelb.edu.au, yeewei.law@unisa.edu.au, {sarah.erfani,caleckie,palani}@unimelb.edu.au

***Abstract*—The ubiquity of mobile sensing devices in the Internet of Things (IoT) enables an emerging data crowdsourcing paradigm called participatory sensing, where multiple individuals collect data and use a cloud service to analyse the union of the collected data. An example of such collaborative analysis is collaborative anomaly detection. Given the possibility that the cloud service is honest but curious, a major challenge is how to protect the participants' privacy. The scheme called Random Multiparty Perturbation (RMP) addresses this challenge by allowing each participant to perturb his/her tabular data by passing the data through a nonlinear function, and projecting the data to a lower dimension using a participant-specific random matrix. Here, we propose an improvement to RMP by introducing a new nonlinear function. The improved scheme is assessed in terms of its recovery resistance to the maximum a priori (MAP) estimation attack. Experimental results and preliminary theoretical analysis indicate that RMP is resistant to collusion attacks and has better recovery resistance to MAP estimation attacks compared to the original scheme. It also achieves a good trade-off between accuracy and privacy.**

I. Introduction

Central to the IoT is the ability to capture data, either through sensors or active contribution by people, which gives birth to the data crowdsourcing paradigm called *participatory sensing*. In this collaborative learning scenario, participants want to extract information/knowledge from their joint records, without disclosing their privacy-sensitive information. The problem of preserving privacy in this scenario is referred to as "privacy-preserving collaborative learning" [1]. Historically, Agrawal and Srikant [2] were the first to study this problem, under the umbrella term "privacy-preserving data mining" (PPDM).

Most PPDM schemes are based on either Secure Multiparty Computation (SMC) or randomisation/perturbation. SMC uses cryptographic primitives to ensure a high level of privacy and accuracy, at the expense of high computational and communication overhead. The main challenge facing SMC-based schemes is the requirement for simultaneous coordination of all participants during the entire training process, which limits the number of participants. A promising alternative to SMC is randomisation, which trades off privacy and accuracy for scalability by perturbing the data in a way that is (i) computationally efficient, (ii) does not allow an attacker to recover the original data, and (iii) does not severely affect the accuracy of data mining. In such a scheme, the participants perturb their records before sending them to the cloud service to be processed. Most existing randomisation approaches such as [1], [3], [4] require all participants to perturb their data using the same perturbation matrix, and are thus vulnerable to collusion (between some of the participants and the cloud service). To address this challenge, Erfani et al. [5] (co-authors of this paper) proposed a privacy-preserving collaborative anomaly detection scheme called *Random Multiparty Perturbation* (RMP) that allows participants to use their own unique, randomly generated perturbation matrix to randomise their data. The perturbation process of RMP consists of a nonlinear transformation stage and a linear projection stage. The nonlinear stage is used to condition the probability density function (pdf) of the perturbed data to thwart *maximum a posteriori* (MAP) estimation attacks, whereas the linear stage is to compress the data and resist independent component analysis attacks. The nonlinear transformation function of RMP is the *double logistic* function, but the privacy-preserving properties of this function have not been thoroughly assessed.

Our contributions here are two-fold: (i) the proposal of an alternative nonlinear transformation function, which we call "repeated Gompertz", and (ii) the evaluation of the privacy-preserving properties of both the double logistic function and the repeated Gompertz function. In terms of the recovery rate, which measures how well an attacker can recover the original data from the perturbed data, the repeated Gompertz function proposed here is found to be more resistant to MAP estimation attacks than the double logistic function proposed in [5].

II. Related work

In general, PPDM schemes are either *syntactic* or *semantic*. Semantic approaches—of which RMP is an example–aim to satisfy some semantic privacy criteria/definitions, which are concerned with minimising the difference between adversarial prior knowledge and adversarial posterior knowledge about the individuals represented in the database. Potentially the most popular semantic privacy criterion is *differential privacy*. Differential privacy was designed for the scenario where a database server *answers queries* in a privacy-preserving manner by adding tailored Laplace noise to the query results [6], [7]. In such a scenario, the database comprises private data of *multiple individuals*. The participatory sensing scenario, where participants are data owners who *publish data* (instead of answering queries) about *themselves alone*, can be considered as a distributed version of the differential privacy scenario. In such a scenario, additional mechanisms need to be paired with differential privacy; evidence supporting this claim can be found in many highly cited references including:

- Shi et al.'s scheme [8] enables participants to upload encrypted values to a data aggregator, which computes the sum of the encrypted values. These values are perturbed with Laplace noise that satisfies (ϵ, δ)-differential privacy, but the encryption relies on a trusted dealer allocating $q+1$ secrets that sum to 0, to the data aggregator and the q participants.
- Ács et al.'s scheme [9] enables smart meters, organised into clusters, to send Laplace noise-tainted readings to an electricity distributor; but requires all meters in a cluster to share pairwise keys.

To dispense with the additional, high-overhead cryptographic mechanisms, most randomisation-based schemes use alternative privacy criteria. For RMP, the criterion *recovery resistance*, defined in Sect. IV, is used. This criterion is based on the *recovery rate* metric used in Sang et al.'s innovative study of attacks on randomisation-based schemes [10].

Randomisation techniques include (i) additive perturbation, (ii) multiplicative perturbation, and (iii) geometric perturbation, and (iv) nonlinear transformation.

Additive perturbation adds independent and identically distributed (i.i.d.) noise to the original data [2], but this additive noise can be filtered out [11].

Multiplicative perturbation premultiplies the original data with a random noise matrix. The following designs of the noise matrix are known:

- *Rotation perturbation* defines the noise matrix as a matrix with orthonormal rows and columns [12]. This scheme is vulnerable to "known-input attacks" [13], where an attacker can recover the original data from its perturbed version with just a few leaked inputs.
- *Random projection* leverages the Johnson-Lindenstrauss Lemma by defining the noise matrix as a matrix whose elements are independently sampled from the same zero-mean Gaussian distribution [3]. If the original data follows a multivariate Gaussian distribution, a large portion of the data can be reconstructed via MAP estimation [14]. Both rotation perturbation and random projection are distance-preserving transformations, which are good for preserving data mining accuracy, but susceptible to attacks that exploit distance relationships [13].
- *Uniform random transformation* (abbreviated as RT) defines the noise matrix as a matrix whose elements are independently sampled from the same uniform distribution [15]. This has the advantage of making attacks on distance-preserving transformations [13] not applicable. RMP uses exactly RT, where the noise matrix is a projection matrix whose elements are independently sampled from the uniform distribution $U(0, 1)$.

Geometric perturbation uses a mix of additive and multiplicative perturbations, where the data matrix $\mathbf{X}$ is mapped to $\mathbf{RX} + \mathbf{\Phi} + \mathbf{\Delta}$, where $\mathbf{R}$ is a rotation perturbation matrix, $\mathbf{\Phi}$ is a random translation matrix with identical entries, and $\mathbf{\Delta}$ is an i.i.d. Gaussian noise matrix [12]. It is known that without $\mathbf{\Delta}$, geometric perturbation is vulnerable to "known input attacks" [13], but there are no general results on how the $\mathbf{\Delta}$ term influences the effectiveness of these attacks. All the randomisation techniques discussed so far are linear techniques.

Nonlinear transformation is meant to be used in conjunction with linear techniques to thwart Bayesian estimation attacks. The general randomisation takes the form $\mathbf{B}+\mathbf{Q}\cdot N(\mathbf{A}+\mathbf{RX})$, where $\mathbf{B}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{R}$ are random matrices, and N is a bounded nonlinear function [4]. The $tanh$ function is found to preserve the distance between normal data points, but collapse the distance between outliers, making the function suitable for privacy-preserving anomaly detection [4], provided only the privacy of anomalous records needs to be protected.

Recently proposed, RMP [5] uses both RT and nonlinear transformation. The innovations of RMP include (i) the use of participant-specific RT matrices, and (ii) the use of the *double logistic* function as the nonlinear transformation function for protecting both anomalous and normal records from Bayesian estimation attacks. However, the privacy-preserving properties of the double logistic function have not been thoroughly assessed. Here, we propose the *repeated Gompertz* function as an alternative to the double logistic function, and provide empirical evidence of the advantages of this alternative in terms of the improved recovery resistance of RMP.

III. RMP: THE IMPROVED SCHEME

As depicted in Fig. 1, the general participatory sensing architecture comprises a set of participants $\mathcal{C} = \{c_i | i = 1, \ldots, q\}$, a data mining cloud service $\mathcal{S}$, and a set of end-users $\mathcal{U}$. The cloud service is assumed to be honest but curious, i.e., it will never perform any malicious action to disrupt the protocols or compromise the participants but it might try to discover privacy-sensitive information of the participants, including colluding with some of the participants. RMP considers the case where (i) the data mining operation is anomaly detection, and (ii) scalability requirements necessitates the use of a randomisation-based PPDM scheme. Based on the state of the art in PPDM, the following design criteria are applicable:

- **Resilience to distance inference attacks**: Uniform random transformation [15] does not preserve the angle (inner product) or Euclidean distance between transformed data points, and is thus resistant to distance inference attacks. Moreover, it is suitable for anomaly detection.
- **Resilience to Bayesian estimation attacks**: Bayesian estimation is a general attack that exploits the pdf of the original data. Gaussian data is particularly exploitable because it reduces the MAP estimation problem to a simple convex optimisation problem [14]. A nonlinear transformation can be applied to prevent this reduction by conditioning the pdf.

- **Resilience to collusion**: Let $\mathbf{X}_i \in \mathbb{R}^{n \times m_i}$ be the nth dimensional dataset of participant c_i, where m_i is the number of records. If participant c_i perturbs its records as $\mathbf{Z}_i = \mathbf{T}\mathbf{X}_i$, and $\mathbf{T} \in \mathbb{R}^{w \times n}$, $w < n$, is a random matrix shared by all participants, then leakage of $\mathbf{T}$ due to collusion to the cloud service S can compromise the privacy of all the participants. Preventing collusion requires each participant to use an independently unique perturbation matrix.

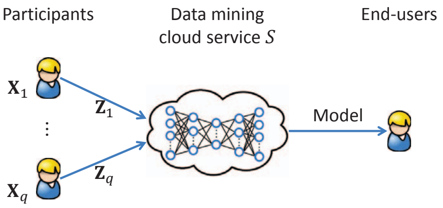

Fig. 1. The general participatory sensing architecture.

A. The improved scheme

RMP's two-stage data perturbation scheme was designed with the preceding criteria in mind. Let $\mathbf{T}$ be a $w \times n$ matrix ($w < n$) with $U(0,1)$-distributed elements. Each participant c_i generates a unique perturbation matrix

$$\tilde{\mathbf{T}}_i = \mathbf{T} + \Delta_i, \tag{1}$$

where each element of Δ_i is drawn from $U(-\alpha, \alpha)$, and $0 < \alpha < 1$. Experimental results show that for small values of α, the accuracy loss in anomaly detection is small. Suppose participant c_i is contributing data $\mathbf{X}_i \in \mathbb{R}^{n \times m_i}$ to the cloud service S for anomaly detection. The participant transforms $\mathbf{X}_i$ to $\mathbf{Z}_i \in \mathbb{R}^{w \times m_i}$ in two stages:

Stage 1: The participant transforms $\mathbf{X}_i$ to $\mathbf{Y}_i$, by applying the nonlinear perturbation function N element-wise:

$$\mathbf{Y}_i = N(\mathbf{X}_i). \tag{2}$$

In the original version of RMP, N is defined as the double logistic function. Here for the improved version of RMP, N is chosen to be the repeated Gompertz function:

$$\begin{aligned} N(x) \overset{\text{def}}{=} & a_1 e^{-b_1 e^{-c_1 x - d_1}} u(0.35 - x) \\ & + \left(0.5 + a_2 e^{-b_2 e^{-c_2 x - d_2}}\right) u(x - 0.35), \end{aligned} \tag{3}$$

where the parameters $a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2$ are defined in Fig. 2, and $u()$ is the Heaviside step function. The derivation of the function parameters is explained in Sect. IV. Fig. 2 plots different nonlinear perturbation functions for comparison.

Stage 2: Using $\tilde{\mathbf{T}}_i$ generated earlier, the participant transforms $\mathbf{Y}_i$ to $\mathbf{Z}_i$:

$$\mathbf{Z}_i = \tilde{\mathbf{T}}_i \mathbf{Y}_i. \tag{4}$$

The participant then sends $\mathbf{Z}_i$ to the cloud service $\mathcal{S}$. Once $\mathcal{S}$ receives all the perturbed datasets $\mathbf{Z}_i$, $i = 1, \ldots, q$, it concatenates them as: $\mathbf{Z}_{\text{all}} = [\mathbf{Z}_1 | \cdots | \mathbf{Z}_q]$, and then trains an anomaly detection model on $\mathbf{Z}_{\text{all}}$. The learned model $\mathcal{M}$ can be used by end-users to identify anomalies in their test records. RMP is independent of the anomaly detection algorithm used, but the autoencoder is used for our study.

Param	Value
a_1	0.5247
b_1	6.0000
c_1	13.2000
d_1	0.2040
a_2	0.4000
b_2	6.0000
c_2	27.5000
d_2	-16.0750

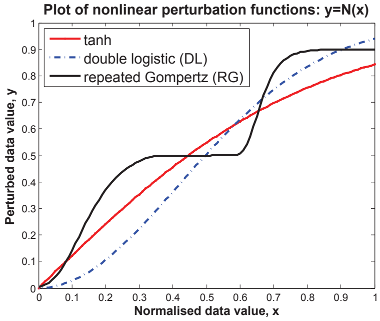

Fig. 2. Parameters of the repeated Gompertz function, and a plot of different nonlinear perturbation functions. The tanh function is $N(x) = \tanh(\beta_t x)$, where $\beta_t \approx 1.23$ [5]. The double logistic function is $N(x) = 1 - \exp(-\beta_{dl} x^2)$, where $\beta_{dl} \approx 2.81$ [5]. The repeated Gompertz function is defined as per Eq. (3).

IV. Privacy analysis

Corresponding to multiplicative perturbation, we use an alternative definition for differential privacy, which we state informally now but formally later: a perturbation scheme is privacy-preserving with respect to an attack $\mathcal{A}$ and a data distribution p_D if only a small fraction of the original data, characterised by p_D, can be recovered from the perturbed data through $\mathcal{A}$. This definition has three major components:

- the *reference attack*;
- the *data distribution*, which captures an aspect of the attacker's auxiliary information; and
- the *recovery rate*, which captures the notions of "small fraction" and "recovered".

To specify the reference attack, we first consider attacks to linear multiplicative perturbation schemes. These types of schemes project a data vector (and hence the whole data matrix) to a lower dimensional space so that an attacker has only an ill-posed problem in the form of an underdetermined system of linear equations $\mathbf{T}\boldsymbol{y} = \boldsymbol{z}$ to work with, where $\boldsymbol{z}$ is a projection of vector $\boldsymbol{y}$. An underdetermined system cannot be solved for $\boldsymbol{y}$ exactly, but given sufficient prior information about $\boldsymbol{y}$, an approximation of the true $\boldsymbol{y}$ may be attainable. We can characterise an attack by the extent of prior information available to the attacker.

In a *known input-output attack*, the attacker has some input samples (i.e., some samples of the original data) and all output samples (i.e., all samples of the perturbed data), and knows which input sample corresponds to which output sample [13]. In the participatory sensing scenario where the cloud service may collude with one or more participants to unravel other participants' data, the known input-output attack is an immediate concern. In the following, our privacy analysis is conducted with respect to a known input-output attack based on MAP estimation — this is our reference attack. MAP estimation is based on Bayesian statistics and is more general than maximum likelihood estimation because the former takes a prior distribution into account.

To measure the strength of the reference attack, we define the recovery rate. If for a data vector $\boldsymbol{x}$ the recovered copy is $\hat{\boldsymbol{x}}$, then the *relative error* is $\xi \overset{\text{def}}{=} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 / \|\boldsymbol{x}\|_2$, where $\|\cdot\|_2$ is the Euclidean norm. Denote the joint distribution of ξ and $\boldsymbol{x}$ by $p_{\Xi,\boldsymbol{X}}(\xi, \boldsymbol{x})$, then we define the *$\epsilon$-recovery rate* with respect to the perturbation algorithm and attack as

$$r_\epsilon(\mathcal{A}, p_D) \overset{\text{def}}{=} \int_{\xi=0}^{\epsilon} \int_{\boldsymbol{x}\in D_{\boldsymbol{x}}} p_{\Xi,\boldsymbol{X}}(\xi, \boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\xi, \tag{5}$$

where $D_{\boldsymbol{x}}$ is the domain of the data vector, and $\boldsymbol{x}$ is normalised. The joint distribution $p_{\Xi,\boldsymbol{X}}$ depends on the attack $\mathcal{A}$ and data distribution p_D. In the absence of an analytical expression for Eq. (5), we estimate the recovery rate as the fraction of test data that can be recovered to within a relative error of ϵ. At this point, we state the privacy definition formally as follows.

A probabilistic algorithm that takes p_D-distributed $\boldsymbol{x} \in \mathbb{R}^n$ as input and produces $\boldsymbol{z} \in \mathbb{R}^w$ as output is *(ϵ, δ)-recovery resistant* with respect to p_D and attack algorithm $\mathcal{A}$ if $r_\epsilon(\mathcal{A}, p_D) = \delta$.

Suppose the attacker is targeting a particular participant by trying to solve $\mathbf{Z} = \mathbf{T}\mathbf{Y}$ for $\mathbf{Y}$. We consider two scenarios: where $\mathbf{T}$ is known, and where $\mathbf{T}$ is unknown.

A. Scenario Where $\mathbf{T}$ is Known

This is the worst-case scenario and here, we assume the attacker somehow knows $\mathbf{T}$ exactly but not $\mathbf{Y}$, for example when the attacker manages to predict the output of the victim's improperly initialised pseudorandom number generator (in fact, such a vulnerability was discovered on the Android mobile platform in mid-2013). Let $\boldsymbol{z}$ represent a column of $\mathbf{Z}$, and $\boldsymbol{y}$ represent a column of $\mathbf{Y}$. The MAP estimate of $\boldsymbol{y}$, given $\mathbf{T}$ and $\boldsymbol{z}$, is

$$\begin{aligned}\hat{\boldsymbol{y}} &= \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{z}, \mathbf{T}) = \arg\max_{\boldsymbol{y}} \frac{p(\boldsymbol{z}|\mathbf{T}, \boldsymbol{y})p(\mathbf{T})p(\boldsymbol{y})}{p(\boldsymbol{z}|\mathbf{T})p(\mathbf{T})} \\ &= \arg\max_{\boldsymbol{y}\in\mathcal{Y}} \frac{p(\boldsymbol{y})}{\int_{\mathbb{R}^n} p(\boldsymbol{z}|\mathbf{T}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{y}} = \arg\max_{\boldsymbol{y}\in\mathcal{Y}} p(\boldsymbol{y}),\end{aligned} \tag{6}$$

where $\mathcal{Y} = \{\boldsymbol{y} : \boldsymbol{z} = \mathbf{T}\boldsymbol{y}\}$.

If $\boldsymbol{y}$ is n-variate Gaussian with a positive definite covariance matrix, then Eq. (6) becomes an easily solvable quadratic programming problem [14, Theorem 1]. The key is to design a nonlinear function N that transforms a potentially Gaussian data distribution to a distribution that deters accurate solution of Eq. (6).

In this paper, we propose using the "repeated Gompertz" function defined in Eq. (3) as the nonlinear function. The following explains how the proposed function is derived. The Gompertz function takes the standard form:

$$\mathrm{Gompertz}(x) = ae^{-be^{-cx}}, \tag{7}$$

where the parameter a specifies the upper asymptote, b controls the displacement along the x axis, and c adjusts the growth rate of the function. As $\tanh(\beta_t x)$ is good for protecting anomalous data points, the repeated Gompertz function is given slopes that approximate those of $\tanh(\beta_t x)$ at $x = 0$ and $x = 1$. The repeated Gompertz function is also designed to have a flat middle section so that for that section the function cannot be inverted, in order to protect normal data points. Through extensive search, we found the geometry in Fig. 2 to be good for protecting both anomalous and normal data points: (i) a Gompertz curve presenting a steep slope over the interval $[0, 0.35]$; and (ii) another Gompertz curve presenting a plateau over the interval $[0.35, 0.6]$, a steeper slope over the interval $[0.6, 0.75]$ and another plateau over the interval $[0.75, 1]$. The parameters of the two Gompertz functions are given in Fig. 2. This compositional structure inspired the name "repeated Gompertz".

B. Scenario where $\mathbf{T}$ is unknown

Consider the case where the attacker knows neither $\mathbf{T}$ nor $\mathbf{Y}$. The MAP estimates of $\mathbf{Y}$ and $\mathbf{T}$, given $\mathbf{Z}$, are

$$\begin{aligned}(\hat{\mathbf{T}}, \hat{\mathbf{Y}}) &= \arg\max_{\mathbf{T},\mathbf{Y}} p(\mathbf{T}, \mathbf{Y}|\mathbf{Z}) \\ &= \arg\max_{\mathbf{T},\mathbf{Y}} \frac{p(\mathbf{Z}|\mathbf{T}, \mathbf{Y})p(\mathbf{T})p(\mathbf{Y})}{\int\int p(\mathbf{Z}|\mathbf{T}, \mathbf{Y})p(\mathbf{T})p(\mathbf{Y})d\mathbf{T}d\mathbf{Y}} \\ &= \arg\max_{(\mathbf{T},\mathbf{Y})\in\Theta} p(\mathbf{T})p(\mathbf{Y}),\end{aligned} \tag{8}$$

where $\Theta \in \{(\mathbf{T}, \mathbf{Y}) : \mathbf{Z} = \mathbf{T}\mathbf{Y}\}$. In a known input-output attack, $p(\mathbf{T})$ and $p(\mathbf{Y})$ are estimated as inputs to Eq. (8). Eq. (8) is a nonconvex optimisation problem that is harder to solve than Eq. (6). The repeated Gompertz is designed to make data recovery via Eq. (6) difficult when $\mathbf{T}$ is known. Now that $\mathbf{T}$ is unknown, the attacker is expected to get an even lower recovery rate by solving Eq. (8), which is a more difficult problem.

V. Simulations and Evaluation

This section presents the simulation and evaluation results of the improved RMP in terms of its privacy-preserving and accuracy-preserving properties.

Experiments are conducted on (i) purely Gaussian datasets, (ii) purely Laplace-distributed datasets, (iii) seven real datasets from the UCI Machine Learning Repository, and (iv) two challenge synthetic datasets. The seven real datasets are (i) Abalone, (ii) Forest, (iii) Adult, (iv) Gas, (v) OAR, (vi) DSA, and (vii) HAR, with dimensionalities of 8, 54, 123, 128, 110, 315 and 561 respectively. The two synthetic datasets are (i) Smiley with 20 features, and (ii) GME with 100 features. The Smiley dataset consists of samples drawn from two compact Gaussians and an arc shaped distribution to resemble a smiley face, and is often used to challenge anomaly detection algorithms. The GME dataset is a mixture of four separated Gaussians.

A. Privacy evaluation

Experimental results are provided in this section on the recovery resistance of the improved RMP against the MAP estimation attack, in terms of the ϵ-recovery rate defined in Eq. (5). In the absence of an analytical expression for Eq. (5), we estimate the ϵ-recovery rate as the fraction of test data that can be recovered to within a relative error of ϵ:

$$\hat{r}_{\epsilon}(\mathcal{A}, p_D) \overset{\text{def}}{=} \frac{\#\left\{\hat{\boldsymbol{x}}_i : \frac{\|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\|_2}{\|\boldsymbol{x}_i\|_2} \leq \epsilon, i = 1, \ldots, m\right\}}{m}, \quad (9)$$

where $\boldsymbol{x}_i$ and $\hat{\boldsymbol{x}}_i$ are the ith original data record and its attacker-estimated value respectively.

To execute MAP estimation, the attacker can either apply the [14, Theorem 1] formula, provided the original data is multivariate Gaussian distributed; or solve the constrained optimisation problem (6). To solve optimisation problem (6), the attacker needs to evaluate an objective function that is the pdf of the original data; for this, the attacker can estimate the pdf of the original data as the pdf of the leaked input samples, using multivariate *kernel density estimation* (KDE). For KDE, we use Ihler and Mandel's Kernel Density Estimation Toolbox for MATLAB[1]. Among the kernels supported, we use the Epanechnikov kernel — which is optimal in the sense of the asymptotic mean integrated squared error — with uniform weights.

TABLE I
EVALUATED SCHEMES

Scheme	Nonlinear perturbation function (stage 1)	Linear projection matrix (stage 2)
RP [3]	none	$\mathbf{T} \sim N_{w \times n}(0, 4)$
tanh+RT	tanh [4]	$\mathbf{T} \sim U_{w \times n}(0, 1)$
DL+RT [5]	double logistic	$\mathbf{T} \sim U_{w \times n}(0, 1)$
RG+RT	repeated Gompertz	$\mathbf{T} \sim U_{w \times n}(0, 1)$

The four schemes shown in Table I are evaluated in the *worst-case* scenario where the attacker knows exactly the victim's perturbation matrix.

Purely Gaussian datasets: Fig. 3 shows that RG+RT provides significantly higher recovery resistance for both normal and anomalous data compared to the other schemes.

Purely Laplace datasets: Fig. 4 shows that RG+RT significantly outperforms other methods for Laplace datasets, and this is especially evident for normal data. Furthermore, the 0.1-recovery rate against RG+RT is below 10%, which is much lower than other schemes.

Assorted real and synthetic datasets: Consistent with the results for purely Gaussian and purely Laplace datasets, as shown in Fig. 5, RG+RT also outperforms tanh+RT and DL+RT in terms of recovery resistance for both normal data and anomalous data. Note the low recovery rates in many cases, especially for example, RG+RT achieves (0.1, 0)-recovery resistance for the DSA and HAR datasets.

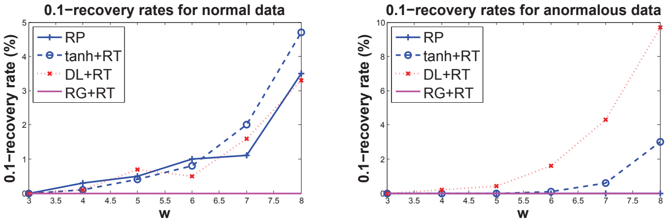

Fig. 3. Recovery rates of MAP estimation attacks against the evaluated schemes, on $w \times 1000$ data projected from 15×1000 normalised Gaussian-distributed data (zero mean, identity covariance matrix).

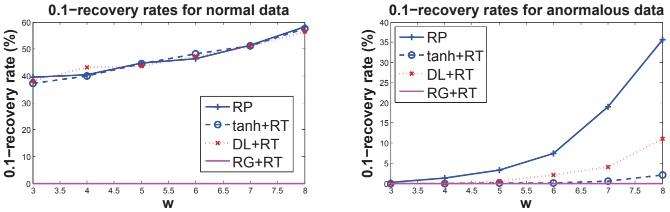

Fig. 4. Recovery rates of MAP estimation attacks against the evaluated schemes, on $w \times 1000$ data projected from 15×1000 normalised Laplace-distributed data (zero mean, unity scale).

B. Accuracy evaluation

For anomaly detection, a stacked denoising autoencoder is used. The hyperparameters of the autoencoder are set based on the best performance on validation set. Feature values in each dataset are normalised to $[0, 1]$ and merged with 5% anomalous records, which are distributed between $[0, 0.05]$ or $[0.95, 1]$. Anomalies are identified by the denoising autoencoder based on the mean absolute error between the inputs and outputs of the training records. According to three sigma rule, a well-known measure for anomaly detection, the reconstruction error is expected to be Gaussian distributed, hence 99.73% of the error values are expected to be at most three standard deviations away from the mean, namely, within the threshold $\mu(e) + 3\sigma(e)$. An error value larger than the threshold is unlikely and is identified as an anomaly.

The Area Under the ROC Curve (AUC) is used to compare the anomaly detection accuracy of the autoencoder with and without data perturbation by our scheme. Without data perturbation, the AUC is close to 1. With data perturbation, the AUC is expected to decrease, and the goal is to measure the extent of this decrement. Reducing data dimensionality from n to $w \leq (n+1)/2$ ensures that no *linear* filter can recover the original data from its perturbed version [3]. On the other hand, this raises the concern of accuracy loss. Fig. 6 shows that dimensionality reduction has minor impact on accuracy. Reducing data dimensionality by 50% decreases the detection rate by at most 5% in the worst case. Our study also reveals that RG+RT has better or similar AUC performance to tanh+RT and DL+RT.

VI. Conclusion and Future Work

We present an improvement to an existing privacy-preserving collaborative anomaly detection scheme called RMP. The randomisation-based scheme perturbs data in two stages: the first, nonlinear stage thwarts Bayesian estimation attacks, whereas the second, linear stage resists independent component analysis, distance inference attacks and collusion attacks. For the nonlinear perturbation stage, a new nonlinear function called the "repeated Gompertz" function is proposed

[1] http://www.ics.uci.edu/~ihler/code/kde.html

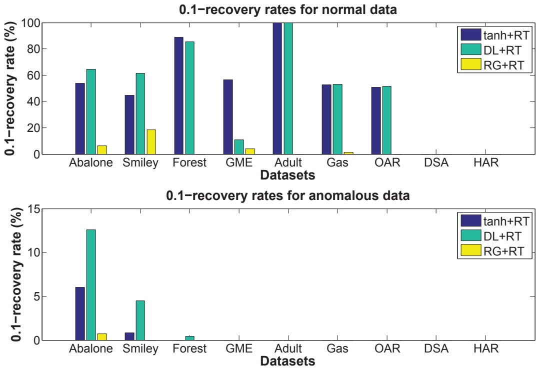

Fig. 5. 0.1-recovery rates of the MAP estimation attack against the evaluated schemes, on various datasets. The rank of the perturbation matrix, w, is set as $\lfloor (n+1)/2 \rfloor$, where n is the number of features. Note zero recovery rates in many cases.

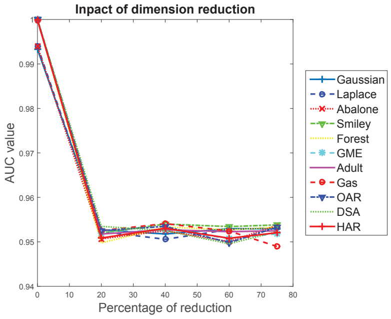

Fig. 6. Impact of dimensionality reduction on the AUC. The x-axis shows the percent reduction in the number of dimensions, $(n-w)/n \times 100\%$.

here. The function is designed to condition the pdf of the perturbed data to protect both anomalous and normal data records. Preliminary analysis and empirical evaluation indicate that the two-stage transformation, RG+RT, maintains privacy of both normal and anomalous data, and delivers the lowest recovery rates for all the selected datasets, outperforming the state of the art (tanh+RT and DL+RT). It also achieves a good trade-off between accuracy and privacy.

The next step from this work is to establish a theoretical framework for designing the nonlinear perturbation function, and evaluating the recovery resistance analytically. There is also a need to extend RMP to other data types and data mining algorithms.

Acknolwedgement

This work is partly supported by the EC under contract CNECT-ICT-609112 (SOCIOTAL).

References

[1] B. Liu, Y. Jiang, F. Sha, and R. Govindan, "Cloud-enabled privacy-preserving collaborative learning for mobile sensing," in *SenSys'12*. ACM, 2012, pp. 57–70.

[2] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *ACM Sigmod Record*, vol. 29, no. 2. ACM, 2000, pp. 439–450.

[3] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 1, pp. 92–106, 2006.

[4] K. Bhaduri, M. D. Stefanski, and A. N. Srivastava, "Privacy-preserving outlier detection through random nonlinear data distortion," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 41, no. 1, pp. 260–272, 2011.

[5] S. M. Erfani, Y. W. Law, S. Karunasekera, C. A. Leckie, and M. Palaniswami, "Privacy-preserving collaborative anomaly detection for participatory sensing," in *Advances in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2014, vol. 8443, pp. 581–593.

[6] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2006, vol. 4052, pp. 1–12.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, 2006, vol. 3876, pp. 265–284.

[8] E. Shi, T.-H. H. Chan, E. G. Rieffel, R. Chow, and D. Song, "Privacy-preserving aggregation of time-series data," in *NDSS*, vol. 2, no. 3, 2011, p. 4.

[9] G. Ács and C. Castelluccia, "I Have a DREAM! (DiffeRentially privatE smArt Metering)," in *Information Hiding*, 2011, vol. 6958, pp. 118–132.

[10] Y. Sang, H. Shen, and H. Tian, "Effective reconstruction of data perturbed by random projections," *IEEE Transactions on Computers*, vol. 61, no. 1, pp. 101–117, 2012.

[11] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in *Proc. 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005, pp. 37–48.

[12] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation," in *Data Mining, Fifth IEEE International Conference on*. IEEE, 2005, pp. 4–pp.

[13] C. R. Giannella, K. Liu, and H. Kargupta, "Breaching euclidean distance-preserving data perturbation using few known inputs," *Data & Knowledge Engineering*, vol. 83, pp. 93–110, 2013.

[14] Y. Sang, H. Shen, and H. Tian, "Effective reconstruction of data perturbed by random projections," *Computers, IEEE Transactions on*, vol. 61, no. 1, pp. 101–117, 2012.

[15] O. L. Mangasarian and E. W. Wild, "Privacy-preserving classification of horizontally partitioned data via random kernels," in *Proc. International Conference on Data Mining (DMIN)*, vol. 2, 2008, pp. 473–479.