Privacy-Preserving Collaborative Anomaly Detection for Participatory Sensing

Sarah M. Erfani[1], Yee Wei Law[2], Shanika Karunasekera[1], Christopher A. Leckie[1], and Marimuthu Palaniswami[2]

[1] NICTA Victoria Research Laboratory, Department of Computing and Information Systems
[2] Department of Electrical and Electronic Engineering, The University of Melbourne, Australia
{sarah.erfani,ywlaw,karusg,caleckie,palani}@unimelb.edu.au

Abstract. In collaborative anomaly detection, multiple data sources submit their data to an on-line service, in order to detect anomalies with respect to the wider population. A major challenge is how to achieve reasonable detection accuracy without disclosing the actual values of the participants' data. We propose a lightweight and scalable privacy-preserving collaborative anomaly detection scheme called Random Multiparty Perturbation (RMP), which uses a combination of nonlinear and participant-specific linear perturbation. Each participant uses an individually perturbed uniformly distributed random matrix, in contrast to existing approaches that use a common random matrix. A privacy analysis is given for Bayesian Estimation and Independent Component Analysis attacks. Experimental results on real and synthetic datasets using an auto-encoder show that RMP yields comparable results to non-privacy preserving anomaly detection.

Keywords: Privacy-preserving data mining, Anomaly detection, Collaborative learning, Participatory sensing, Horizontally partitioned data.

1 Introduction

Anomaly detection (also known as outlier detection) plays a key role in data mining for detecting unusual patterns or events in an unsupervised manner. In particular, there is growing interest in *collaborative anomaly detection* [1,2,3], where multiple data sources submit their data to an on-line service, in order to detect anomalies with respect to the wider population. For example, in *participatory sensing networks* (PSNs) [4], participants collect and upload their data to a central service to detect unusual events, such as the emergence of a source of pollution in environmental sensing, or disease outbreaks in public health monitoring. A major challenge for collaborative anomaly detection in this context is how to maintain the trust of participants in terms of both the accuracy of the anomaly detection service as well as the privacy of the participants' data. In this paper, we propose a random perturbation scheme for privacy-preserving

V.S. Tseng et al. (Eds.): PAKDD 2014, Part I, LNAI 8443, pp. 581–593, 2014.

anomaly detection, which is resilient to a variety of privacy attacks while achieving comparable accuracy to anomaly detection on the original unperturbed data.

There have been several studies on collaborative anomaly detection, where a number of participants want to build a global model from their local records, while none of the participants are willing to disclose their private data. Most existing work relies on the use of Secure Multiparty Computation (SMC) to generate the global model for anomaly detection [1,2]. While this approach can achieve high levels of privacy and accuracy, SMC incurs a high communication and computational overhead. Moreover, SMC based methods require the simultaneous coordination of all participants during the entire training process, which limits the number of participants. Thus, an open research challenge is how to improve scalability while achieving high levels of accuracy and privacy.

To address this challenge, we propose a privacy-preserving scheme for anomaly detection called *Random Multiparty Perturbation* (RMP). RMP supports the scenario where participants contribute their local data to a public service that trains an anomaly detection model from the combined data. This model can then be distributed to end-users who want to test for anomalies in their local data. In this paper, we focus on the use of an auto-associative neural network (AANN, also known as an autoencoder) as our anomaly detection model, although our scheme is also applicable to other types of anomaly detectors.

In order for the participants of RMP to maintain the privacy of their data, we propose a form of random perturbation to be used by each participant. Previous approaches to random perturbation in this context [5,6,7,3] require all participants to perturb their data in the same way, which makes this scheme potentially vulnerable to breaches of privacy if collusion occurs between rogue participants and the server. In contrast, RMP proposes a scheme in which each participant first perturbs their contributed data using a unique, private random perturbation matrix. In addition, any end-user can apply the resulting anomaly detection model to their own local data by using a public perturbation matrix. This provides a scalable collaborative approach to anomaly detection, which ensures a high level of privacy while still achieving a high level of accuracy to the end-users.

The main contributions of this paper are as follows: (i) We propose a privacy-preserving model of collaborative anomaly detection based on random perturbation, such that each participant in training the anomaly detector can have their own unique perturbation matrix, while the resulting anomaly detection model can be used by any number of users for testing. (ii) In contrast to previous methods, we give the first privacy-preserving scheme for auto-associator anomaly detectors that does not require computationally intensive cryptographic techniques. (iii) We show analytically the resilience of our scheme to two types of attacks—Bayesian Estimation and Independent Component Analysis (ICA). (iv) We demonstrate that the accuracy of our approach is comparable to non-privacy preserving anomaly detection on various datasets.

2 Related Work

In general, privacy-preserving data mining (PPDM) schemes can be classified as *syntactic* or *semantic*. Syntactic approaches aim to prevent syntactic attacks, such as the table linkage attack. Semantic approaches aim to satisfy semantic privacy criteria, which are concerned with minimising the difference between adversarial prior knowledge and adversarial posterior knowledge about the individuals represented in the database. While both approaches have the same goal, i.e., hiding the real values from the data miner, syntactic approaches typically take the entire database as input, whereas semantic approaches do not. Since in participatory sensing the participants are responsible for masking their own data before outsouring them to the data miner, semantic approaches are the only option here.

In the following we review existing work on privacy-preserving back-propagation and anomaly detection in the context of collaborative learning, and highlight the major differences with our work. The majority of semantic PPDM approaches, either in the context of privacy-preserving back-propagation [8,9] or privacy-preserving anomaly detection [1,2], use Secure Multiparty Computation (SMC). SMC approaches suffer from high computational complexity. Moreover, they require the cooperation of all participants throughout the whole process of building the data mining models, and hence suffer a lack of scalability.

Another approach to semantic PPDM is data *randomisation*, which refers to the randomised distortion or perturbation of data prior to analysis. Perturbing data elements, for example, by introducing additive or multiplicative noise, allows the server to extract the statistical properties of records without requiring the participants to allocate substantial resources or coordinate with the server during the training process. We now outline the main types of data perturbation methods: additive, multiplicative, and nonlinear transformation.

Additive perturbation adds noise to the data [10,11]. Since independent random noise can be filtered out [12,13], the general principle is that the additive noise should be correlated with the data [14,15].

Multiplicative perturbation multiplies the data with noise. A typical approach is to use a zero-mean Gaussian random matrix as the noise for a distance-preserving transformation [5]. However, if the original data follows a multivariate Gaussian distribution, a large portion of the data can be reconstructed via attacks to distance-preserving transformations as proposed in [16,17].

In general, distance-preserving transformations are good for accuracy but are susceptible to attacks that exploit distance relationships. In comparison, the *random transformation* approach [7], where the noise matrix is a random matrix with elements uniformly distributed between 0 and 1, is not distance-preserving and hence not susceptible to these attacks.

Nonlinear transformations can be used to break the distance-preserving effect of a distance-preserving transformation on some data points. For example, the function tanh approximately preserves the distance between normal data points, but collapses the distance between outliers [3]—it is hence suitable for anomaly detection. In RMP, we use the *double logistic* function for this purpose and for conditioning the probability density function (pdf) of the transformed data.

The aforementioned works satisfy privacy requirements in the case where all participants are assumed to be semi-honest (i.e., are not colluding with other parties) and agree on using the same perturbation matrix. As discussed in Section 3.1, this assumption is restrictive in applications such as participatory sensing. A few works have tried to extend current randomisation approaches to a collaborative learning architecture [6,18]. For example, Chen et al.'s framework [6] is a multiparty collaborative mining scheme using a combination of additive and multiplicative perturbations. This scheme requires the participants to stay in touch for an extended period to generate the perturbation matrix, which is impractical for large-scale participatory sensing. Liu et al. [18] build a framework on top of [5], and allow each participant to perturb the training data with a distinct perturbation matrix. This scheme then regresses these mathematical relationships between training samples in a unique way, thereby preserving classification accuracy. Although participants use private matrices, the underlying transformation scheme is still vulnerable to distance inference attacks [16,17].

In summary, data randomisation is the most promising approach to PPDM in the paradigm of PSN. Unlike distance-preserving transformations, random transformation does not preserve Euclidean distances and inner products between data distances, hence it thwarts distance inference attacks. However, if this method is applied to collaborative learning, all the participants must agree on the same perturbation matrix, and then collusion attacks may succeed. Furthermore, the mining models generated from the perturbed records are specific to the participants. For these reasons our RMP scheme uses a random transformation and provides each participant with a private perturbation matrix. The perturbation matrices are tailored so that both data privacy and accuracy of the mining models are achieved. In addition, RMP generates models that can be adopted by any user, i.e, data contributors and non-contributors.

3 RMP Privacy-Preserving Anomaly Detection Scheme

In this section we present our Random Multiparty Perturbation (RMP) scheme, which considers a general participatory sensing architecture comprising three types of parties, namely, participants, the data mining server and end-users.

3.1 Problem Statement and Design Principles

We consider the case of a participatory sensing network that comprises a set of participants, $\mathcal{C} = \{c_i | i = 1, \ldots, q\}$, a mining server $\mathcal{S}$, and an arbitrary number of end-users $\mathcal{U}$, as can be seen in Fig. 1. Each participant c_i is an individual who captures data records for the same set of attributes, i.e., a horizontal partition, and contributes the sampled records to $\mathcal{S}$ for training purposes. The server $\mathcal{S}$ is a third-party providing a data mining service to the participants. After receiving the contributed data records, the server trains an anomaly detector that generates a global classification model $\mathbf{M}$ from the locally collected data. The end-users, $\mathcal{U}$, could be the participants themselves or third parties such

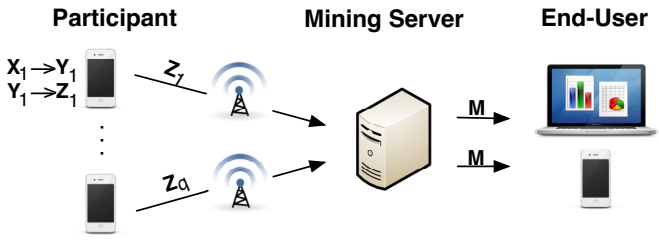

Fig. 1. The RMP architecture

as analysts trying to learn about the monitored phenomena. We share a similar underlying assumption with [4], in which the computational demands on the participants should be minimised, while the anomaly detection model should be in a form that can be disseminated and used by an arbitrary number of end-users, and not limited to the original participants or customised for a small number of end-users. In addition, we adopt a well established assumption in the literature that regards all parties as semi-honest entities, i.e., they follow the protocol. To make the scenario realistic, we assume that a small subset of the participants might collude with the server to infer other participants' submitted records. Designing a collaborative data mining scheme that fulfills these requirements with low communication and computation costs while preserving the participants' privacy is an open problem.

For privacy, we need to ensure that the attribute values in the participants' contributed records are properly masked: given the masked values, the server cannot infer the original values. However, this must be done without oversacrificing accuracy, i.e., the results of anomaly detection based on the masked data should be close to the corresponding result using the original data. Multiplicative perturbation projects data to a lower dimensional space. The perturbed data matrix has a lower rank than the original data matrix, thereby forcing the attacker to solve an undetermined system of linear equations. However, this is not enough, and the following design principles are pertinent.

Resilience to Distance Inference Attacks: The review of existing multiplicative perturbation schemes in Section 2 reveals that distance-preserving transformations are susceptible to distance inference attacks [16,19]. The challenge is to find a non-distance preserving transform that is suitable for certain data mining tasks. *Random transformation* [7] qualifies as such a transform in that it does not preserve the dot product or Euclidean distance among transformed data points, yet 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 a maximum a posteriori estimation problem into a simple convex optimization problem [17]. A suitable defense is to prevent this reduction by conditioning the pdf through a nonlinear transformation.

Resilience to Collusion: Let $\mathbf{X}_i \in \mathbb{R}^{n \times m_i}$ be participant c_i's dataset, and $\mathbf{T} \in \mathbb{R}^{w \times n}$ be a random matrix *shared by all participants*, where $w < n$. The participant c_i perturbs its records as $\mathbf{Z}_i = \mathbf{T}\mathbf{X}_i$. Since the constructed anomaly detector is perturbed, the perturbation matrix needs to be shared with all end-users. This approach of using a common $\mathbf{T}$ poses a serious privacy risk. If a rogue participant or end-user colludes with the server, the server can recover any participant's original data using the breached perturbation matrix. The naive solution of generating an arbitrarily different perturbation matrix $\mathbf{T}_i$ for each participant c_i does not work, because building an accurate mining model requires consistency among the perturbation matrices. To overcome this challenge, RMP generates participant-specific perturbation matrices by perturbing $\mathbf{T}$.

3.2 The Scheme

RMP is designed to address the design principles in the previous subsection. 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 in 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. Next, we describe RMP in full detail.

In RMP, a participant transforms $\mathbf{X}_i$ to $\mathbf{Z}_i$ in two stages:

Stage 1: The participant transforms $\mathbf{X}_i$ to $\mathbf{Y}_i$, by applying a *double logistic function* to $\mathbf{X}_i$ element-wise:

$$y_{k,l} \stackrel{\text{def}}{=} \text{sgn}(x_{k,l})[1 - \exp(-\beta x_{k,l}^2)], \tag{2}$$

for $k = 1, \ldots, n$, $l = 1, \ldots, m_i$, where sgn is the signum function. For suitable values of β and normalised values of $x_{k,l}$ (i.e., $|x_{k,l}| \leq 1$), the double logistic function approximates the identity function $y_{k,l} = x_{k,l}$ well. To maximise this approximation, we equate the optimal value of β to the value that minimises the integral of the squared difference between the two functions:

$$\beta^\star = \arg\min_{\beta} \int_0^1 [1 - \exp(-\beta x^2) - x]^2 dx \approx 2.81.$$

The role of this nonlinear transformation is to condition the pdf of $\mathbf{Y}_i$ to thwart Bayesian Estimation attacks, as detailed in Section 5.

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

$$\mathbf{Z}_i \stackrel{\text{def}}{=} \tilde{\mathbf{T}}_i \mathbf{Y}_i. \tag{3}$$

The participant then sends $\mathbf{Z}_i$ to the mining server $\mathcal{S}$, which will receive the following from all the participants:

$$\mathbf{Z}_{\text{all}} \stackrel{\text{def}}{=} [\mathbf{Z}_1 | \mathbf{Z}_2 | \cdots | \mathbf{Z}_q] .$$

The role of $\mathcal{S}$ is to learn an anomaly detection model encoding the underlying distribution of $\mathbf{Z}_{\text{all}}$. End-users given access to the model and $\mathbf{T}$ can then detect anomalies in their data with respect to the model. RMP is independent of the anomaly detection algorithm used, but the auto-associative neural network is used for our study and is discussed next.

4 Anomaly Detection Service

In our collaborative framework, the role of the server $\mathcal{S}$ is to learn an anomaly detection function, which can then be disseminated for use by the end-users. In particular, the learned anomaly detection function should encode a model of the underlying distribution of the (perturbed) training data provided by the participants. A given data record can then be tested using this model for anomalies.

While our general framework can potentially use a wide variety of anomaly detection models, in this paper we use an auto-associative neural network (AANN) [20], also known as an auto-encoder, as the basis for our anomaly detection function. We choose to use an AANN for our anomaly detection function because: (i) it can be trained in an unsupervised manner on either normal data alone, or a mixture of normal data with a small but unspecified proportion of anomalous data; (ii) it is capable of learning a wide variety of underlying distributions of training data; and (iii) the resulting anomaly detection function is compact and computationally efficient—hence practical for dissemination to end-users.

An AANN is a multi-layer feed-forward neural network that has the same number of output nodes as input nodes. Between the input and output layers, a hidden "bottleneck" layer of a smaller dimension than the input/output layers captures significant features in the inputs through compression. Training the AANN means adjusting the weights of the network for each training record, so that the outputs closely match the inputs. In this way, the AANN learns a nonlinear manifold that represents the underlying distribution of the data. In a trained AANN, the reconstruction error (integrated squared error between the inputs and outputs) should be high for anomalies, but low for normal data. Let e_i be the reconstruction error for training dataset X_i. If the reconstruction error for a test record is larger than the threshold $\theta = \mu(e_i) + 3\sigma(e_i)$, where $i = 1, \ldots, q$, the record is identified as an anomaly. Due to space constraints, the interested reader is referred to [20] for the details of the AANN training algorithm.

5 Privacy Analysis

In the *statistical disclosure control* and *privacy-preserving data publishing* literature, the most popular semantic privacy criterion is *differential privacy*. Differential privacy is for *answering queries* to a database containing private data of *multiple individuals*. For the participatory sensing scenario where participants are data owners who *publish data* (instead of answering queries) about *themselves alone*, differential privacy is not a good fit. Below, we propose an alternative (informal) privacy criterion.

Linear multiplicative perturbation schemes aim to project a 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 $\tilde{\mathbf{T}}\boldsymbol{x} = \boldsymbol{y}$ to work with, where $\boldsymbol{y}$ is a projection of data vector $\boldsymbol{x}$ and the projection matrix $\tilde{\mathbf{T}}$ is assumed known in the worst case. An underdetermined system cannot be solved for $\boldsymbol{x}$ exactly, but given sufficient prior information about $\boldsymbol{x}$, an approximation of the true $\boldsymbol{x}$ might be attainable. In a *known input-output attack*, the attacker has some input samples (i.e., some samples of $\boldsymbol{x}$) and all output samples (i.e., all samples of $\boldsymbol{y}$), and knows which input sample corresponds to which output sample [21,22,19]. In the collaborative learning scenario where the data miner 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 two known input-output attacks, one based on Bayesian estimation, and one based on Independent Component Analysis (ICA). Suppose the attacker is targeting a particular participant by trying to solve $\mathbf{Z} = \tilde{\mathbf{T}}\mathbf{Y} = (\mathbf{T} + \Delta)\mathbf{Y}$ for $\mathbf{Y}$. In the analysis below, let $\boldsymbol{z} \sim \mathcal{Z}$ represent a column of $\mathbf{Z}$, and $\boldsymbol{y} \sim \mathcal{Y}$ represent a column of $\mathbf{Y}$.

5.1 Attacks Based on Bayesian Estimation

We consider two scenarios where $\tilde{\mathbf{T}}$ is known and where $\tilde{\mathbf{T}}$ is unknown.

Scenario where $\tilde{\mathbf{T}}$ is Known: For a worst-case analysis, we assume the attacker somehow knows $\tilde{\mathbf{T}}$ exactly but not $\mathbf{X}$. In a Bayesian formulation, the *maximum a posteriori* (MAP) estimate of $\boldsymbol{y}$, given $\tilde{\mathbf{T}}$ and $\boldsymbol{z}$, is

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p_{\mathcal{Y}}(\boldsymbol{y}|\boldsymbol{z}, \tilde{\mathbf{T}}) = \arg\max_{\boldsymbol{y}\in\mathcal{Y}} p_{\mathcal{Y}}(\boldsymbol{y}), \tag{4}$$

where $\mathcal{Y} = \{\boldsymbol{y} : \boldsymbol{z} = \tilde{\mathbf{T}}\boldsymbol{y}\}$ [17]. Note that MAP estimation is a more general approach than maximum likelihood estimation because the former takes a prior distribution (which in our case is $p_{\mathcal{Y}}$) into account. If $p_{\mathcal{Y}}$ is an n-variate Gaussian with a positive definite covariance matrix, then (4) becomes a quadratic programming problem with solution [17, Theorem 1]:

$$\hat{\boldsymbol{y}} = \bar{\boldsymbol{y}} + \boldsymbol{\Sigma}_{\mathcal{Y}}\tilde{\mathbf{T}}'\boldsymbol{\Sigma}_{\mathcal{Z}}^{-1}(\boldsymbol{z} - \tilde{\mathbf{T}}\bar{\boldsymbol{y}}), \tag{5}$$

where $\bar{\boldsymbol{y}}$ and $\boldsymbol{\Sigma}_{\mathcal{Y}}$ are the sample mean vector and sample covariance matrix of $\mathcal{Y}$ respectively, and $\boldsymbol{\Sigma}_{\mathcal{Z}}$ is the sample covariance matrix of $\mathcal{Z}$. Note that $\boldsymbol{\Sigma}_{\mathcal{Y}}$ *is* positive definite, provided the covariance matrix is full rank and there are more samples of $\mathcal{Y}$ than dimensions of $\mathcal{Y}$ [23]. In this case, we can write $\boldsymbol{\Sigma}_{\mathcal{Y}} = \mathbf{Q}\mathbf{Q}'$, where $\mathbf{Q}$ is a nonsingular matrix. Furthermore, $\boldsymbol{\Sigma}_{\mathcal{Z}} = \tilde{\mathbf{T}}\boldsymbol{\Sigma}_{\mathcal{Y}}\tilde{\mathbf{T}}' = \tilde{\mathbf{T}}\mathbf{Q}\mathbf{Q}'\tilde{\mathbf{T}}'$. Since $\tilde{\mathbf{T}}$ is nonsingular, $\boldsymbol{\Sigma}_{\mathcal{Z}}^{-1}$ and therefore the solution (5) exists.

The analysis above suggests that to thwart MAP estimation, we cannot hope to generate $\tilde{\mathbf{T}}$ such that $\boldsymbol{\Sigma}_{\mathcal{Z}}$ is singular. Instead, it is more productive to prevent MAP estimation from being reducible to a quadratic programming problem solvable by (5) in the first place. RMP achieves this by nonlinearly transforming

the original data. If X is a standard Gaussian random variable, then based on [24, Section 4.7], the pdf of $Y = \text{sgn}(X)[1 - \exp(-\beta X^2)]$ is

$$p_Y(y) = \frac{1}{\sqrt{8\pi\beta \ln\left(\frac{1}{1-|y|}\right)}} \left(\frac{1}{1-|y|}\right)^{1-\frac{1}{2\beta}}, |y| < 1, y \neq 0.$$

Unlike a standard Gaussian which is continuous everywhere and has a global maximum at $x = 0$, $p_Y \to \infty$ at $y = -1, 0, 1$. Using p_Y or $\ln p_Y$ in (4) renders the optimisation problem non-convex, and numerically unstable near $y = -1, 0, 1$, which is problematic for numerical methods such as hill climbing and simulated annealing. Applying the same nonlinear transformation to Laplace-distributed data (sparse data) has the same effect. Therefore, RMP's nonlinear transformation converts a potentially Gaussian (Laplace) data distribution to a non-Gaussian (non-Laplace) one that hampers the attacker's solution of (4).

The double logistic function is better than tanh, which is used in [3], in terms of thwarting MAP estimation attacks. If X is a standard Gaussian random variable, then the pdf of $Y = \tanh(X)$ is $p_Y(y) = \frac{1}{(1-y^2)\sqrt{2\pi}} \exp(-\frac{\text{artanh}(y)^2}{2}), |y| < 1$. For $|y| \leq 0.9$, the pdf above is convex. This means the attacker can solve (4) as a convex optimisation problem, when the data are perturbed using tanh.

Scenario Where $\tilde{\mathbf{T}}$ is Unknown: In the previous scenario, the attacker knows $\mathbf{T}$ and the relationship between $\mathbf{T}$ and $\tilde{\mathbf{T}}$ (see Equation 1). We also consider the scenario where the attacker does not know $\tilde{\mathbf{T}}$. Note that even with precise knowledge of $\mathbf{T}$ and α, without further information, any matrix value between $\mathbf{T} - \alpha\mathbf{1}$ and $\mathbf{T} + \alpha\mathbf{1}$ can be an estimate of the victim's matrix $\tilde{\mathbf{T}}$. According to Lemma 1, for every element of $\tilde{\mathbf{T}}$, there is a 50% chance of guessing its value wrong by at least $(2 - \sqrt{2})\alpha$.

Lemma 1. *Let D be the difference between two $U(-\alpha, \alpha)$-distributed random variables. Then for $0 \leq d \leq 2\alpha$, $\Pr[|D| \geq d] = (2\alpha - d)^2/(4\alpha^2)$.*

Proof. Let A and B be two $U(-\alpha, \alpha)$-distributed random variables. Then the pdf of $D = A - B$ is given by the convolution

$$p_D(d) = \frac{1}{2\alpha} \int_{-\alpha}^{\alpha} p_{U(-\alpha,\alpha)}(d + b)\mathrm{d}b = \begin{cases} \frac{1}{2\alpha}\left(1 + \frac{d}{2\alpha}\right) & -2\alpha \leq x < 0, \\ \frac{1}{2\alpha}\left(1 - \frac{d}{2\alpha}\right) & 0 \leq x < 2\alpha, \\ 0 & \text{elsewhere.} \end{cases}$$

To get $\Pr[|D| \geq d]$ where $0 \leq d \leq 2\alpha$, we integrate the expression above from -2α to $-d$, and from d to 2α. With a little algebra we get $\Pr[|D| \geq d] = \frac{(2\alpha-d)^2}{4\alpha^2}$.

MAP estimation can be used to estimate both $\mathbf{Y}$ and $\tilde{\mathbf{T}}$. The MAP estimates of $\mathbf{Y}$ and $\tilde{\mathbf{T}}$, given $\mathbf{Z}$, are the matrix values that maximise $p_{\tilde{\mathcal{T}},\mathcal{Y}|\mathcal{Z}}(\tilde{\mathbf{T}}, \mathbf{Y}|\mathbf{Z}) = p_{\tilde{\mathcal{T}}}(\tilde{\mathbf{T}}|\mathbf{Z}) p_{\mathcal{Y}}(\mathbf{Y}|\mathbf{Z})$. The optimisation problem can be written as

$$\max_{\tilde{\mathbf{T}},\mathbf{Y}} \; p_{\tilde{\mathcal{T}}}(\tilde{\mathbf{T}}) p_{\mathcal{Y}}(\mathbf{Y}) \text{ s.t. } \mathbf{Z} = \tilde{\mathbf{T}}\mathbf{Y};$$

or the following when $p_{\tilde{\mathcal{T}}}$ is substituted with $U_{w\times n}(\mathbf{T} - \alpha\mathbf{1}, \mathbf{T} + \alpha\mathbf{1})$ and $p_{\mathcal{Y}}$ is assumed to be zero-mean Gaussian:

$$\min_{\tilde{\mathbf{T}},\mathbf{Y}} \sum_{j=1}^{m} \boldsymbol{y}_j' \Sigma_{\mathcal{Y}}^{-1} \boldsymbol{y}_j \text{ s.t. } \mathbf{Z} = \tilde{\mathbf{T}}\mathbf{Y}, \mathbf{T} - \alpha\mathbf{1} \preceq \tilde{\mathbf{T}} \preceq \mathbf{T} + \alpha\mathbf{1}. \tag{6}$$

In (6), $\boldsymbol{y}_j$ $(j = 1, \ldots, m)$ are columns of $\mathbf{Y}$. Note that in the equality constraint, both $\tilde{\mathbf{T}}$ and $\mathbf{Y}$ are optimisation variables, so even the Gaussian assumption does not reduce (6) to a convex problem. As previously explained, RMP's nonlinear transformation converts a potentially Gaussian (Laplace) data distribution to a non-Gaussian (non-Laplace) one. This hampers the attacker's solution of not only (4) but also (6), which is a harder problem than (4).

5.2 Attacks Based on Independent Component Analysis (ICA)

ICA can be used to estimate $\tilde{\mathbf{T}} \in \mathbb{R}^{w\times n}$ and $\mathbf{Y} \in \mathbb{R}^{n\times m}$, knowing only their product $\mathbf{Z} = \tilde{\mathbf{T}}\mathbf{Y}$, provided (i) $w = n$ (the even-determined case) or $w > n$ (the over-determined case); (ii) the attributes (rows of $\mathbf{Y}$) are pairwise independent; (iii) at most one of the attributes are Gaussian; (iv) $\tilde{\mathbf{T}}$ has full column rank.

However, RMP enforces $w < n$. When $w < n$, the problem of ICA becomes *overcomplete ICA* (or underdetermined ICA). In this case, the mixing matrix $\tilde{\mathbf{T}}$ is identifiable but the independent components are not [22]. Furthermore, when $w \leq (n+1)/2$, no linear filter can separate the observed mixture $\mathbf{Z}$ into two or more disjoint groups, i.e., recover any row of $\mathbf{Y}$ [5].

6 Empirical Results

In this section we evaluate the quality of our proposed privacy-preserving anomaly detection scheme RMP when used with an AANN. The main objective of our experiment is to measure the trade-off in accuracy of our AANN anomaly detection algorithm as a result of maintaining the participants' privacy. Note that Lemma 1 provides a theoretical relationship between α and the privacy that can be achieved in terms of an attacker's ability to estimate the value of a target victim's perturbation matrix $\tilde{\mathbf{T}}$. Thus, in our empirical evaluation, we consider the effect of different levels of privacy in terms of α on the overall accuracy of anomaly detection. We use the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) to measure the performance of RMP. The effectiveness and change in accuracy of RMP are evaluated by comparing against a non-privacy-preserving neural network, in which the raw data records are fed to the AANN for training and testing.

Experiments are conducted on four real datasets from the UCI Machine Learning Repository (all collected from sensor networks except the fourth): (i) Human Activity Recognition using Smartphones (HARS), (ii) Opportunity activity

Table 1. Comparing AUC values of RMP against non-privacy anomaly detection

			α				
Dataset	n	Raw	0.01	0.1	0.2	0.4	0.6
Abalone	8	1	0.95	0.92	0.87	N/A	N/A
Banana	8	1	0.98	0.92	0.81	N/A	N/A
Gas	168	0.99	0.98	0.98	0.98	0.95	0.92
HARS	561	0.99	0.98	0.98	0.97	0.96	0.95
Opportunity	242	0.99	0.99	0.98	0.94	0.90	0.87

Note: "Raw" indicates the non-privacy preserving anomaly detection on the non-perturbed data, and α values indicate the level of imposed noise in RMP anomaly detection.

recognition (Opportunity), (iii) Gas Sensor Array Drift (Gas), (iv) Abalone. We also use the Banana synthetic dataset, generated from a mixture of overlapping Gaussians to resemble a pair of bananas in any two dimensions. We ran the experiment on the first 1000 records of each dataset. Feature values in each dataset are normalised between $[0, 1]$ and merged with 5% anomalous records, which are randomly drawn from $U(0, 1)$. In each experiment a random subset of the dataset is partitioned horizontally among the participants in batches of 30 records and submitted to the server for training.

We deploy a three-layer AANN with the same number of input and output units, i.e., the number of units is set corresponding to the dataset attributes n. The number of hidden units for each dataset is set to about half of the input units (empirically we found that increasing the number of hidden units causes overfitting). All weights and biases in the neural network are initialised randomly in the range of $[-0.1, 0.1]$. The learning rate and momentum are set to 0.25 and 0.85, respectively, and the number of training epochs range from 300 to 1000.

Table 1 compares the results using an AANN anomaly detector on the unperturbed data records ("Raw") along with the corresponding results using the privacy preserving scheme of RMP. Accuracy in RMP is affected by its two stages of transformation, i.e., applying the double logistic function and the random transformation, and the level of added noise α. As can be seen from the table, when data are perturbed with a marginal level of noise $\alpha = 0.01$, the accuracy decreases slightly (about 1%). Hence, it shows that the transformations do not exert a significant impact on the accuracy of anomaly detection. In the reported

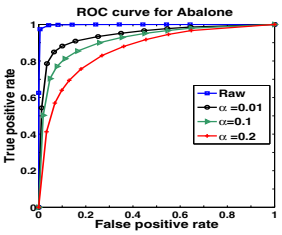

Fig. 2. Abalone Dataset

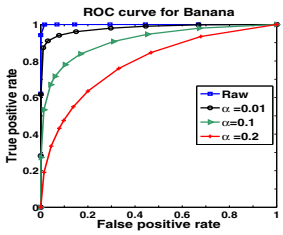

Fig. 3. Banana Dataset

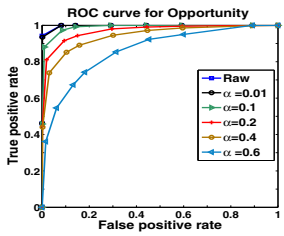

Fig. 4. Opportunity Dataset

results in Table 1, the dimensionality of the datasets is reduced by $r = 1$, where $r = n - w$. Our empirical experiments show that the accuracy on datasets with a larger number of attributes n are less affected by an increase of r, e.g., reducing n by 40% only decreases accuracy by about 1%. However, that is not the case for datasets with small n, e.g., the Abalone and Banana datasets, where reducing the data dimensionality by half might result in a 10% reduction in accuracy.

The level of added noise to the perturbation matrices has a major influence on RMP accuracy. As α increases, so does the loss in accuracy, especially in datasets with smaller numbers of attributes. Since the accuracy loss is generally small, RMP is a highly effective approach for privacy-preserving anomaly detection.

7 Conclusion and Future Work

In a typical participatory sensing scenario, participants send data to a data mining server; the server builds a model of the data; and end-users download the model for their own analyses. Collaborative anomaly detection refers to the case where the model is for anomaly detection. RMP is a privacy-preserving collaborative learning scheme that masks the participants' data using a combination of nonlinear and linear perturbations, while maintaining detection accuracy. RMP protects the private data of participants using individual perturbation matrices, imposes minimal communication and computation overhead on the participants, and scales for an arbitrary number of participants or end-users. We show analytically how RMP is resilient to two common types of attacks: Bayesian Estimation and ICA. Our experiments show that RMP yields comparable results to non-privacy preserving anomaly detection using AANN on a variety of real and synthetic benchmark datasets. As follow-up to this preliminary work, we are in the process of establishing a mathematical framework that relates α to accuracy loss and privacy level—this will allow us to determine α based on the intended trade-off between accuracy and privacy. We are also investigating the resilience of RMP, in conjunction with other supervised or unsupervised learning algorithms, to attacks exploiting sparse datasets, such as overcomplete ICA.

Acknowledgments. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. Yee Wei Law is partly supported by ARC DP1095452 and the EC under contract CNECT-ICT-609112 (SOCIOTAL).

References

1. Dung, L.T., Bao, H.T.: A Distributed Solution for Privacy Preserving Outlier Detection. In: Third International Conference on KSE, pp. 26–31 (2011)
2. Vaidya, J., Clifton, C.: Privacy-preserving outlier detection. In: IEEE ICDM, pp. 233–240 (2004)
3. Bhaduri, K., Stefanski, M.D., Srivastava, A.N.: Privacy-preserving outlier detection through random nonlinear data distortion. IEEE TSMC, Part B 41, 260–272 (2011)

4. Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., Srivastava, M.B.: Participatory sensing. In: ACM SenSys 1st Workshop on World-Sensor-Web: Mobile Device Centric Sensor Networks and Applications (2006)
5. Liu, K., Kargupta, H., Ryan, J.: Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE TKDE 18(1), 92–106 (2006)
6. Chen, K., Liu, L.: Privacy-preserving multiparty collaborative mining with geometric data perturbation. IEEE TPDS 20(12), 1764–1776 (2009)
7. Mangasarian, O.L., Wild, E.W.: Privacy-preserving classification of horizontally partitioned data via random kernels. In: Proceedings of DMIN (2007)
8. Bansal, A., Chen, T., Zhong, S.: Privacy preserving Back-propagation neural network learning over arbitrarily partitioned data. Neural Computing and Applications 20, 143–150 (2010)
9. Zhang, Y., Zhong, S.: A privacy-preserving algorithm for distributed training of neural network ensembles. Neural Computing and Applications 22, 269–282 (2012)
10. Agrawal, R., Srikant, R.: Privacy-preserving data mining. ACM SIGMOD Record 29(2), 439–450 (2000)
11. Agrawal, D., Aggarwal, C.C.: On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of ACM PODS, pp. 247–255 (2001)
12. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy preserving properties of random data perturbation technique. In: IEEE ICDM, pp. 99–106 (2003)
13. Huang, Z., Du, W., Chen, B.: Deriving private information from randomized data. In: Proceedings of ACM SIGMOD, pp. 37–48 (2005)
14. Papadimitriou, S., Li, F., Kollios, G., Yu, P.S.: Time series compressibility and privacy. In: Proceedings of VLDB, pp. 459–470 (2007)
15. Ganti, R.K., Pham, N., Tsai, Y.E., Abdelzaher, T.F.: Poolview: stream privacy for grassroots participatory sensing. In: Proceedings of ACM SenSys, pp. 281–294 (2008)
16. Liu, K., Giannella, C., Kargupta, H.: A survey of attack techniques on privacy-preserving data perturbation methods. In: Privacy-Preserving Data Mining. Springer (2008)
17. Sang, Y., Shen, H., Tian, H.: Effective reconstruction of data perturbed by random projections. IEEE Transactions on Computers 61(1), 101–117 (2012)
18. Liu, B., Jiang, Y., Sha, F., Govindan, R.: Cloud-enabled privacy-preserving collaborative learning for mobile sensing. In: Proceedings of SenSys, pp. 57–70 (2012)
19. Giannella, C.R., Liu, K., Kargupta, H.: Breaching euclidean distance-preserving data perturbation using few known inputs. Data & Knowledge Engineering (2012)
20. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature 323(6088), 533–536 (1986)
21. Liu, K., Giannella, C., Kargupta, H.: An attacker's view of distance preserving maps for privacy preserving data mining. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 297–308. Springer, Heidelberg (2006)
22. Guo, S., Wu, X.: Deriving private information from arbitrarily projected data. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 84–95. Springer, Heidelberg (2007)
23. Dykstra, R.L.: Establishing the positive definiteness of the sample covariance matrix. The Annals of Mathematical Statistics 41(6), 2153–2154 (1970)
24. Grimmett, G.R., Stirzaker, D.R.: Probability and Random Processes, 3rd edn. Oxford University Press (2001)