Random Projection to Preserve Patient Privacy

Aris Anagnostopoulos
Sapienza University of Rome
aris@diag.uniroma1.it

Fabio Angeletti
Sapienza University of Rome
angeletti@diag.uniroma1.it

Federico Arcangeli
Sapienza University of Rome
federico.arcangeli1@gmail.com

Chris Schwiegelshohn
Sapienza University of Rome
schwiegelshohn@diag.uniroma1.it

Andrea Vitaletti
Sapienza University of Rome
vitaletti@diag.uniroma1.it

Abstract

With the availability of accessible and widely used cloud services, it is natural that large components of healthcare systems migrate to them; for example, patient databases can be stored and processed in the cloud. Such cloud services provide enhanced flexibility and additional gains, such as availability, ease of data share, and so on. This trend poses serious threats regarding the privacy of the patients and the trust that an individual must put into the healthcare system itself. Thus, there is a strong need of privacy preservation, achieved through a variety of different approaches.

In this paper, we study the application of a *random projection*-based approach to patient data as a means to achieve two goals: (1) provably mask the identity of users under some adversarial-attack settings, (2) preserve enough information to allow for aggregate data analysis and application of machine-learning techniques. As far as we know, such approaches have not been applied and tested on medical data. We analyze the trade-off between the loss of accuracy on the outcome of machine-learning algorithms and the resilience against an adversary. We show that random projections proved to be strong against known input/output attacks while offering high quality data, as long as the projected space is smaller than the original space, and as long as the amount of leaked data available to the adversary is limited.

1 Introduction

During the recent years, we witnessed the tremendous progress made in the field of wireless sensor networks. This paved the way and facilitated the wide adoption of small electronic devices with interconnection capabilities. These devices composed the majority of the so-called Internet of Things (IoT). The IoT is a highly dynamic and radically distributed networked system, composed of an incredible high number of objects [MSDPC12]. It is vastly considered as one of the most expanding area within future technologies and it is attracting vast attention in different industry applications [LL15], ranging from smart cities to home automation, farming, and many more fields of application. Ubiquitous sensors, smart objects and devices involved in IoT can generate a tremendous amount of data [JDXC+14]. This flow of data requires robust, available and fast storage solutions and builds the bases to very effective and powerful algorithms in the fields of machine learning and data mining [CDW+15].

Electronics health-care solutions and, generally speaking, the Internet of Health Things (IoHT) follows the same trend. About 73% of healthcare executives say that IoHT is a disrupting techology for the next years and it is becoming one of the most funded areas in IoT. Pervasive IoHT enables cost savings for both the administrations and the individuals, but on the other hand it has some barriers, like: privacy and

security concerns, lack of skilled workers, poor interoperability and more [acc17].

Given the sensitive nature of healthcare data, there is a strong need to protect the information of the patients. Furthermore, the recent adoption of General Data Protection Regulation (GDPR) strengthens data protection and now it must be applied to any organisation or individual that collects and processes information related to EU citizens, regardless where the data is physically stored or where they are based[Alb16, Tan16]. At the same time, analysis of such data are crucial for medical research and the drug industry. Consequently, there is a need to design approaches that allow data processing without exposing the personal underlying information.

For this reason, there has been a series of techniques for perturbing data such that information on individual data points cannot be leaked, while aggregate information is preserved. Examples of such approaches are k-anonymity [SS98] and differential privacy [Dwo06]. The various approaches put different importance on the privacy requirements; for instance, differential privacy attempts to alter the data such as to provide very strong privacy guarantees, typically, without specifying the usefulness of the resulting data. for general-purpose data analysis.

In this paper we apply a method, which can be found in [Liu07a], where the explicit goal is to obtain a dataset that remains useful after the perturbation (still providing some privacy guarantees). More specifically, our approach is based on *random projections* (RP), a technique that is typically applied primarily for efficiency reasons. It is based on a fundamental result from the work of Johnson and Lindenstrauss [JL84] and the idea is the following: Assume that we have large amounts of data (n datapoints), lying on a high-dimensional space. Then, if we project each point to a random subspace of dimension $O(\log n/\epsilon^2)$, with high probability all pairwise distances between the data points are preserved within a factor of $1 \pm \epsilon$. This technique has found multiple applications in streaming algorihtms, in finding nearest neighbors in high-dimensional spaces, in reducing the dimension of databases, and so on.

Our idea of applying a random projection approach on privacy-preserving data mining is the following. Assume that we have the records of multiple patients. Then we can consider a random projection of these data. The result of Johnson and Lindenstrauss guarantees that if we execute algorithms that depend on the pairwise distances on the data (e.g., several clustering or classification algorithms), then the results obtained are with high probability similar to those obtained on the real data set (and the error can be quantified). Furthermore, because the projections are random, one cannot use the projected data to obtain the real data: each datapoint appears to be random. This, unless the attacker has some significant power. This trade off is studied in previous works (e.g. [BBM16], [Liu07a]).

It is not clear a priori that this approach could work in the application on medical dataset that interests us. For instance, the lemma by Johnson and Lindenstrauss is typically applied on settings where the original data lie on a very high-dimensional space. However, in practice, the original dimension may be low (for instance in our dataset it is about 50). In this paper we look at this and other issues by applying the random projection to a dataset containing information about 70K cases of diabetes [SDG+14]. We show that it is possible to reduce the dimensionality of the data and still obtain accuracy scores that are comparable with the ones obtained from the original non-projected data. At the same time, we also show how sensible and private patient information such their age or gender are safe against attacks that try to reconstruct the mapping between the original data and the projected data after applying random projections.

Structure of the paper. The rest of this paper is organized as follows. In Section 2 we present current solutions and the state of the art on random projections and other privacy-preserving techniques. In Section 3 we present the goals of our work and the approach we used for achieving them, leveraging random projections. In Section 4 we show our experimental results, where we explore the limits of our approach both in terms of accuracy and privacy protection. We conclude in Section 5 where we also propose future work.

2 State of the art

Data leakages are very common [Wik14]. In this work, we are more interested in reducing the ability of an attacker to reconstruct non-yet-leaked data from the leaked one. Within medical premises, there are multiple individuals who could obtain access to protected information from the rightous doctor to untrustful workers. This could lead to multiple entities knowing protected personal health information.

Before 2003, with the enforcing of HIPAA rules, some private medical information were regularly shared among professional [Bro00]. Following the guidelines from Health Insurance Portability and Accountability Act (HIPAA) [fDCP+03], the US government made the first concrete attempt to mitigate the chance of re-identification of patients. In 2009, it was clear that HIPAA is not sufficient to protect the privacy of an individual. In fact, the HIPAA was not

able to protect the user personal information after the anonimization process that substituted HIPAA parameters with IDs. In a famous case [New09], some researchers were able to re-identify users and also their sexual orientation and other information. Moreover, the availability of correlated data (coming from the same source or other sources) could greatly help to identify a patient. Data breaches continue to increase year after year, between 2005 and 2014, only in the US, more than 26 million of people had some form of personal health information leaked [Wik14].

Therefore, more elaborate tecniques, which add noise to the data, have been developed in the last years to protect users' privacy and still maintain a good level of accuracy when exploring and analyzing the data. One of these is differential privacy [Dwo06]. This approach focuses on providing statistically coherent responses querying a database, i.e. third parties are interested to query for information about a sample of a population, not a single individual. Instead, we are interested also in providing data about a specific individual, for example investigating if he or she is suitable for a clinical trial.

In [KKMM12a] the authors proved that the Johnson-Lindenstrauss transform can be used as an alternative approach to achieve differential privacy. The method is then compared against other techniques, such as adding Gaussian noise to data or randomized response. The proposed approach has superior accuracy bounds than the others, while still keeping secure the privacy of the records. The authors also criticize the work of Liu et al. about releasing data to third parties after applying random projections in order to protect sensitive information while still preserving accuracy of different data mining algorithms: an adversary that has some background knowledge can infer approximations of the original data. We address this issue in the scenario of known input-output data (section 3.2) and show how in real world scenario regarding medical data, under reasonable assumptions for the power of the adversary, it is difficult for an attacker to discover private information from projected records.

In the literature there exist a very large number of works regarding the re-identification of person starting from various data, within some degree it is called "breaking the k-anonimity". For example in [NS08] the authors presents a method to re-identify a user from its preferences.

In this paper, we aim at investigating to what extent RPs can provide useful data for machine learning algorithms (e.g. classification) on a group of potential patients while preserving at the same time the privacy of individuals. RPs have been employed in a number of healthcare applications, for example to segment tumor areas [KEDR12], to enhance tomography [FMR10], to cluster DNA microarrays[AV09] or to classify cancer [XLZW16].

In [LKR06], RPs were used to mask clear data projecting them in smaller spaces, while in [BBM16] and [KKMM12b], similarly to our work, the authors discuss how to exploit RPs to enhance data privacy. The authors in [LKR06] also discuss the utility of the RP in reducing complexity of problems while maintaining the usefulness of the projected data for algorithms. It is anticipated that by 2020 there will be more than 26 billion devices involved in IoT related applications [RvdM14]. Surely, not all of them will be part of the healthcare field, however we expect a very large amount of information to process. The usefulness of RP in reducing problem complexity (or resource requirements) is well understood and exploited as useful resource in the literature [CEM+15, LKR06, FB03, BZD10, AWY+99]. For example, in [FB03], the authors explore some ways to reduce high dimensional data for clustering while, in [LF12], is presented a work on classification of small patches of images from a very large database that takes advantage of the properties offered by RPs.

During the last two decades, the contribution of machine learning and data mining algorithms in healthcare applications became more frequent year after year. This is well demonstrated in the literature, for example in [CHH+17, MP99, ZWC+09, HHC+14]. One last aspect to consider is the chance to link together multiple datasets. For example in [LJJ+09], the authors presented the infrastructure of a databank in order to enable record-linkage research studies. This linkage on one hand could deeply help the development of newer treatments or drugs, but on the other hand poses threats to the privacy of the individuals.

3 Problem Formulation

We consider a reference scenario in which a group of users, characterized by private features, are potentially suitable for a clinical trial. Only a limited number of users in the group will be actually enrolled in the trial. For the enrolled users, namely the patients, the private features will be eventually made public to participate to the clinical trial in the most effective way. Some knowledge on the group is of primary importance for the researchers to understand the size and the characteristics of potential patients. In general, users are well disposed to support this need of the researchers provided that their privacy is preserved. The main problem we want to address in this paper is:

Can we learn something on the group of users as a whole, while preserving the privacy of the individuals

who will not participate in the trial?

More formally, we consider a group of n users, where each user u is characterized by m features. We represent the corresponding dataset as a matrix $X \in \mathbb{R}^{m \times n}$, with m rows (the features) and n columns (the users). As already observed, in the era of big data, m and n can be particularly big.

Giannella et al. [GLH13] show how it is possible to break the privacy in some contexts of distance preserving mappings. Liu [Liu07b] instead, highlights how mappings that do preserve distances within certain bounds like random projections can boost the privacy guarantees. We will apply these techniques in order to prove that users' privacy can be kept safe against malicious attackers.

We are interested in understanding to what extent the random projection technique, which has been originally conceived to reduce the dimensionality of a dataset, can also be used to preserve the privacy of the users. In particular, we apply a random projection to X, such that if $R \in \mathbb{R}^{k \times m}$ is the random-projection matrix $Y = RX$ is the transformed matrix after applying the random projection, with $Y \in \mathbb{R}^{k \times n}$. We denote by x_i^u the column in X associated to user u_i, and with y_i^u the corresponding column in Y. In the scenario we are describing the projected matrix Y is known to the public, it is indeed the dataset on which researchers try to distill information on the group; the transformation matrix R and the original data X are private. Some columns of X may become public once the corresponding users will eventually decide to participate to a clinical trial, in other words some pairs (x_i^u, y_i^u) will become public.
We can now better describe the problem, splitting it into two sub-problems:

Accuracy. *Can we learn something on the group exploiting Y?* Here we want to understand if the results of some machine-learning algorithms on Y are a good approximation of the ones obtained on X. If we answer positively to this question, we can at least conclude that what can be learned from the original data can be also learned from the projected data.

Privacy. *Can we preserve the privacy of the individuals that will not participate in the trial?* As already observed, Y is public whereas only some columns of X will eventually become public when the corresponding users will decide to participate in a clinical trial. Consequently, some pairs (x_i^u, y_i^u) will become public. Here we want to understand if an attacker knowing Y and the some pairs (x_i^u, y_i^u) can possibly know something about the other users that do not participate in the trial.

We now elaborate on these two dimensions.

3.1 Accuracy

Lemma 3.1 provides a technique to generate a low-dimensional representation of the original data maintaining the pairwise distance within an error ϵ. Since the pairwise distance is the key ingredient for many classification tasks performed by machine learning algorithms, this property allow us to have some guarantees that the solution found in the low-dimensional space is a good approximation of the solution in the original and higher dimensional space. Furthermore, reducing the size of the input data speeds-up the execution time of the algorithms and limits the amount of resources needed.

Lemma 3.1 *(Johnson and Lindenstrauss) Given $\epsilon > 0$ and an integer n let k be a positive integer such that $k \geq k_0 = O(\frac{log(n)}{\epsilon^2})$. For every set P of n points in $\mathbb{R}^m$ there exists a mapping $f : \mathbb{R}^m \to \mathbb{R}^k$ such that for all $u, v \in P$*

$$(1-\epsilon) \parallel u - v \parallel^2 \leq \parallel f(u) - f(v) \parallel^2 \leq (1+\epsilon) \parallel u - v \parallel^2$$

It can be proved that a random projection, is a mapping f that fulfills the previous lemma with positive probability. This is often referred as *JL-embeddings*.

3.2 Privacy: Known Input–Output Attack

We now try to answer one of the questions we raised in the previous section: Can a a malicious third party who knows some pairs (x_i^u, y_i^u) (i.e. that a particular record x_i^u is associated to y_i^u after its projection) learn information about other records?

Liu in his Ph.D. thesis [Liu07b] describes a *Bayes privacy model* to measure the privacy offered by a perturbation technique. The model considers the attacker's apriori and a posteriori beliefs about the data and uses Bayesian inference to evaluate the privacy. For completeness, we repeat his framework here.

Let x be the unknown private data, y the perturbed data and θ the attacker's additional knowledge about the data. Then the MAP estimate of x given y and θ is

$$\hat{x}_{MAP}(y, \theta) = \arg\max_x f_{x|y,\theta}(x|y, \theta)$$

with $f_{x|y,\theta}$ the conditional probability density of x given y and θ.

Let X_p denote the first p columns of X and X_{n-p} the remaining columns. We define similarly Y_p and Y_{n-p}. We further assume that the columns of X_p are linearly independent and that X_p is known to the attacker (i.e., the attacker has full knowledge of p patients). Y is entirely known to the attacker, because

as we stated before, it is publicly available to conduct experiments on the projected data.

For the next reasoning the following hypothesis must be verified:

- The original data arose from as a sample from a matrix variate distribution.
- The projection matrix R is a $k \times m$ random matrix with each entry indipendent and identically distributed with 0 mean and unit variance. R has a matrix variate Gaussian distribution with mean matrix $M = 0$ and covariance matrix $\Sigma = I_k \otimes I_n$.[1]
- Y has a matrix variate Gaussian distribution with mean matrix $M = 0$ and covariance matrix $\Sigma = I_k \otimes \frac{1}{k} X^T X$

The attacker will try to produce $\hat{x_i}$, with $1 \leq i \leq m-p$, such that $\hat{x_i}$ is a good estimate of the undisclosed private record x_i. In other words the attacker's target is to try to give an estimation of one of the records contained in X_{n-p}, given that he knows the records in X_p and their randomly projected counterpart in Y_p.

We now derive the MAP estimate of x given $y = Rx$ and the known matrices X_p and Y_p

$$\hat{x}_{MAP}(y, \theta) = \arg\max_x f_{x|y,\theta}(\mathbf{x} = x|\frac{1}{\sqrt{k}}Rx = y, \frac{1}{\sqrt{k}}RX_p = Y_p)$$

which can be simplified in

$$\arg\max_x f_{x,y,\theta}(\frac{1}{\sqrt{k}}R\overline{X} = \overline{Y})$$

where $\overline{X} = [xX_p]$ and $\overline{Y} = [yY_p]$.

We further suppose that the attacker has no other background knowledge about the private data, so we can assume that $\theta = 0$.

The previous result can be written as

$$\arg\max_x f_{x,y}(\frac{1}{\sqrt{k}}R\overline{X} = \overline{Y}) =$$
$$\arg\max_x f_{\frac{1}{\sqrt{k}}RZ|Z}(\frac{1}{\sqrt{k}}RZ = \overline{Y}|Z = \overline{X}) f_Z(Z = \overline{X})$$

If we assume that f_Z is distributed uniformly over an interval, we finally get

$$\hat{x}_{MAP}(y) = \arg\max_x f_{\frac{1}{\sqrt{k}}RZ|Z}(\frac{1}{\sqrt{k}}RZ = \overline{Y}|Z = \overline{X})$$

In [Liu07b, Theorem 5.3.8] is shown that the probability density function we obtained has the following form

$$(2\pi)^{-\frac{1}{2}k(p+1)} det(\frac{1}{k}\overline{X}^T\overline{X})^{-\frac{1}{2}k} etr\{-\frac{1}{2}\overline{Y}(\frac{1}{k}\overline{X}^T\overline{X})^{-1}\overline{Y}^T\}$$

[1] $\otimes$ indicates the Kroenecker product of two matrices [Liu07b]

We want to maximize this function in order to solve the problem of finding the best estimate of x given the observation of X_p.

Liu proposes an algorithm to estimate the nondisclosed records of a certain dataset. Experimental results have shown that while decreasing the number of column records known to the attacker (denoted by p) the relative error of the estimation increases. The error in the estimation increases also decreasing the dimensionality of the projected subspace (denoted by k). In particular the algorithm uses the Nelder–Mead simplex algorithm to find the optimal solution of the maximization problem.

4 Experimental results

In this section, we present experimental results obtained on a dataset containing information about 70000 cases of diabetes diagnosticated in 130 US hospitals during the decade 1999-2008 [SDG+14] [2]. From now on we will refer to this dataset as the *diabetes dataset*.

We focus on the *classification* of patients based on their privatized data. Following the work in [DMS+17, HXY+16], we choose to use *random forest classifier* in our dataset to classify the users. Moreover, from the work in [Jol17], we know that random forest classifiers works really well with random projections. In Figure 1 we report the effectiveness in terms of accuracy running the random forest classifier [Pal05] on the original data and on the projected data in multiple lower dimensional spaces. To run and validate the classification algorithm, we divided the whole dataset into two parts: *train* and *test*. In the dataset we decided to predict the range of glucose level in the blood. So that, the algorithm was firstly trained with the records within the *train* part of the *diabetes dataset*, providing all the target values. Thus, we made the random forest classifier algorithm predicts the target values in the *test* part giving its features as input. Moreover, we tested the effectiveness of RPs also with *k-nearest neighbors (k-NN)* classifier, the results were reported in Figure 2. Our approach was inspired by [AC06]. The results are quite different because in the first experiment we taken a feature of the dataset (the range of glucose level in the blood) as the value to predict, instead with the second experiment we choose to run firstly a *kMeans* clustering algorithm (on the whole dataset) to obtain labeled groups and then, with the *k-nearest neighbors (k-NN)* classifier we predicted the values.

The blue line represents the accuracy of the machine

[2] The dataset is called "Diabetes 130-US hospitals for years 1999–2008 Data Set" and is available at https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+ye

learning algorithm on the original data. The orange line, instead, represents the accuracy of the same algorithm on the projected (obfuscated) data. We tested the classification algorithm on projected spaces in different sizes, starting from only 2 components up to 10 components.

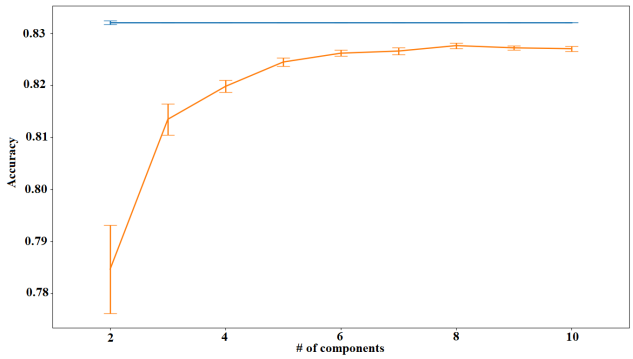

Figure 1: Accuracy of the *random forest classifier* algorithm on the original data (blue line) and on the projected data (orange line), varying the projection space (# of components). Mean values are reported as lines and 95% confidence intervals are reported as vertical lines.

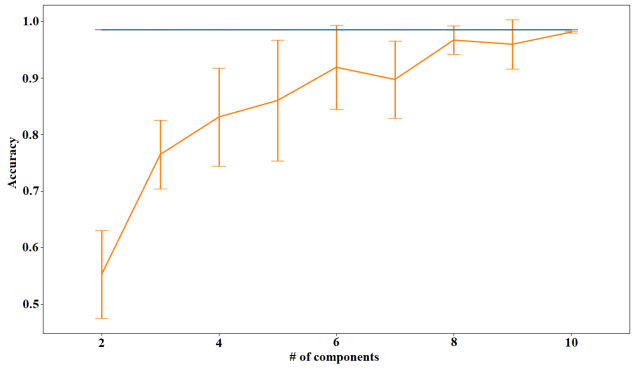

Figure 2: Accuracy of the *k-nearest neighbors (k-NN)* algorithm on the original data (blue line) and on the projected data (orange line), varying the projection space (# of components). Mean values are reported as lines and 95% confidence intervals are reported as vertical lines.

The lines plotted in Figure 1 presents the average values for each projection space, while the vertical wiskers represent the confidence interval corresponding to a specific projection space. For the baseline (classification on the clear data) we ran the classification algorithm 50 times, in each round starting from a random state of the random forest classifier. Since the original data is not projected into any space, we have only a baseline with the associated mean value and confidence interval. Thus, we reported the confidence interval only at the lefties part of line using wisker again. Instead, for the accuracy of classification on the projected data, we ran the algorithm more than 100 times. In each round the algorithm generated a value for each projected space. The results were obtained using the *scikit-learn* package on *Python 3.6*.

In [KLR06, LGK06] the authors explore the security of such techniques: they show how it is possible to use data dimensionality reduction techniques to lower the complexity of data mining algorithms while preserving their accuracy and how those techniques preserve the privacy of users.

The authors start from the same privacy hypothesis we have presented in 3.2 and study how an attacker in possession of a collection of linearly independent private data records and their corresponding transformed part can gather some insight about other records.

We present the results we got running the algorithm of [Liu07b] on this dataset. After choosing a number p of record pairs (x_p, y_p) we select a record x for which we do not know the mapping; the algorithm we are using will try to give an estimation $\hat{x}$ of the original record x.

We used two techniques to evaluate how similar to the original records the algorithm's estimations were. We measured the distance between the estimation $\hat{x}$ provided by the algorithm and the original record x. We compute the relative error between the two vectors with the following:

$$E(x, \hat{x}) = \frac{||x - \hat{x}||_2}{||x||_2}$$

The error E increases with the Euclidean distance between the two. Notice that with this notation it may happen that the error is greater than one: this could verify in the case that the distance between x and its estimations $\hat{x}$ is high and the norm of x is a small value. This could happen if the algorithm's estimation is very far off from the original record.

This measuring has the drawback to lack an upper bound for the dissimilarity. Neither the cosine similarity helps, since in our case we are not interested only in the direction of vectors but also in their magnitude.

A solution is provided in [JNY07], where a radial basis function kernel can be used for representing similarities: we are going to use $1 - \frac{1}{e^{dist(x,\hat{x})}}$ as a similarity function between x and its estimation $\hat{x}$, where $dist(x, \hat{x}) = ||x - \hat{x}||^2$. The bigger the Euclidean distance between two vectors, the bigger the error $e^{dist(x,\hat{x})}$ will be. In this way we have a $[0, 1)$ bound for the similarity of the estimations. By applying the inverse we get a value in the range [0,1): if x and $\hat{x}$ are the same vector (perfect reconstruction performed by the algorithm) then $\frac{1}{e^{dist(x,\hat{x})}} = 1$.

Our workplan is the following: for every subspace of dimensionality k we apply the algorithm with differ-

ent knowledge about the number of pairs (x_p, y_p) the attacker knows. We go from $p = k-1$ to $p = 1$. In the next figures we display the results of our experiments, with the two different measuring techniques we used to quantify the similarity between the original records and the estimated ones. We report the mean of the errors for every pair (k, p) and the variance. On the X axis are placed the tuples (k, p) for which we have conducted the experiments, on the Y axis we placed the reconstruction errors.
On low-dimensionality subspaces we get a high relative error, meaning that it is not possible to give an effective approximation of the original (private) data records. In higher dimensions the approximation is closer to the original data. We ran our experiments with 10 features of the dataset, since with vectors of higher dimensionality it becomes more difficult to run the reconstruction algorithm in reasonable times; also with higher dimensionalities the algorithm we are using outputs vector reconstructions that are very dissimilar from the original ones.

We applied the random projection to reduce the feature space in different dimensions, from 10 to 3. Notice, however, that even when the projected space has the same dimension of the original space, we already get a significant relative error, meaning that on the average it is not possible for the attacker to extrapolate any useful information about the patients' records. So for records of higher dimensionality there is already a safe privacy bound when applying random projection to them, at least against this kind of attacks.

We assigned an increasing numerical value to nominal features, that is, we assigned 0 to the text *male* and 1 to text *female* in the *gender* feature.

We applied random projection to this records, from $k = 10$ (no dimensionality reduction) to $k = 3$; the number p of pairs (original record, projected record) known to the attacker is in the range $k - 1 \leq p \leq 1$.

With $k = 2$ we obviously have only $p = 1$: we omit this result since it is not meaningful with respect the other results we get for higher k and p, because it does not show how knowing less (or more) information about the original data changes the reliability of the reconsturction we get.

In the next figures we show the mean and variances of the errors for every tuple (k, p) for which we have conducted the experiment. It can be seen from the charts that as the number of known input-output pairs p decreases, the reconstruction error increases. Together with the dimensionality reduction, disclosing a scarce number of known input-output pairs can help with the task of preserving the privacy of users involved in clinical trials.

In this case we are projecting low dimensionality vectors ($k = 10$) but we still get high reconstruction errors when applying the techniques we have explained. This is another confirmation of the thesis that random projections help keep the privacy of users when their information is shared among research institutes.

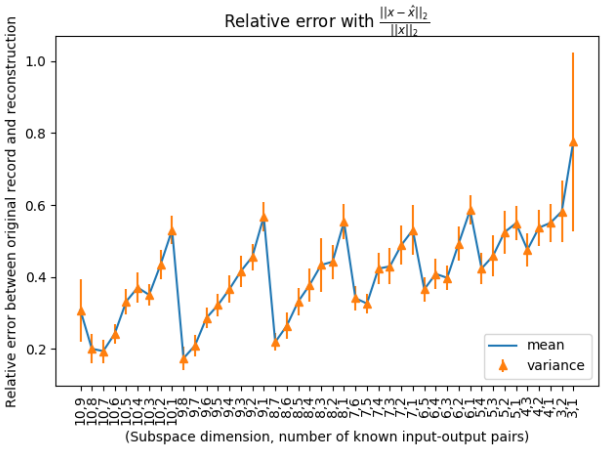

Figure 3: Mean and variance of the relative error while using the formula $\frac{||x-\hat{x}||_2}{||x||_2}$

Figure 4: Mean and variance of the similarity between original records and their reconstruction while using the similarity function $1 - \frac{1}{e^{||x-\hat{x}||_2}}$

5 Conclusions

In this work, we applied an random-projections approach to privacy-preserving data mining of medical data.

First we demonstrated the usefulness of RP in increasing privacy of personal health data. The projected data are useful for machine learning algorithms (for example, in clustering) while allows the sharing of information between parties without revealing the patients' clear data. In this particular application, this is of notable importance since allows entities involved

in different health branches to cooperate effectively without sharing clear data. Second, we investigated to what extent an attacker can discover additional information starting from leaked data. As long as the projected space is smaller than the original space, and as long as the amount of data leaked is small, than the proposed approach is robust and mantains very good performance in both accuracy and privacy.

We analyzed the ratio behind and the performances (in terms of accuracy) of the RP applied on sensible healthcare data. The results shows that the use of RP offers great enhancements in privacy protection. This was a first step into developing a full-fledged platform that allows the effective share of medical data. In future we are planning a bigger real-world deployment of such platform to further validate our results, plus an audit to check privacy protection against real third parties.

References

[AC06] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing, Seattle, WA, USA, May 21-23, 2006*, pages 557–563, 2006.

[acc17] accentureconsulting. Internet of health things survey, 2017.

[Alb16] Jan Philipp Albrecht. How the gdpr will change the world. *Eur. Data Prot. L. Rev.*, 2:287, 2016.

[AV09] Roberto Avogadri and Giorgio Valentini. Fuzzy ensemble clustering based on random projections for dna microarray data analysis. *Artificial Intelligence in Medicine*, 45(2-3):173–183, 2009.

[AWY+99] Charu C Aggarwal, Joel L Wolf, Philip S Yu, Cecilia Procopiuc, and Jong Soo Park. Fast algorithms for projected clustering. In *ACM SIGMoD Record*, volume 28, pages 61–72. ACM, 1999.

[BBM16] Tiziano Bianchi, Valerio Bioglio, and Enrico Magli. Analysis of one-time random projections for privacy preserving compressed sensing. *IEEE Transactions on Information Forensics and Security*, 11(2):313–327, 2016.

[Bro00] Anthony The Guardian Browne. Lives ruined as nhs leaks patients' notes, 2000.

[BZD10] Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems*, pages 298–306, 2010.

[CDW+15] Feng Chen, Pan Deng, Jiafu Wan, Daqiang Zhang, Athanasios V Vasilakos, and Xiaohui Rong. Data mining for the internet of things: literature review and challenges. *International Journal of Distributed Sensor Networks*, 11(8):431047, 2015.

[CEM+15] Michael B Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 163–172. ACM, 2015.

[CHH+17] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5:8869–8879, 2017.

[DMS+17] Sarah DuBrava, Jack Mardekian, Alesia Sadosky, E Jay Bienen, Bruce Parsons, Markay Hopps, and John Markman. Using random forest models to identify correlates of a diabetic peripheral neuropathy diagnosis from electronic health record data. *Pain Medicine*, 18(1):107–115, 2017.

[Dwo06] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, pages 1–12, Berlin, Heidelberg, 2006. Springer-Verlag.

[FB03] Xiaoli Z Fern and Carla E Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 186–193, 2003.

[fDCP+03] Centers for Disease Control, Prevention, et al. Hipaa privacy rule and public health. guidance from cdc and the us department of health and human services. *MMWR: Morbidity and mortality weekly report*, 52(Suppl. 1):1–17, 2003.

[FMR10] Yi Fang, Sundar Murugappan, and Karthik Ramani. Estimating view parameters from random projections for tomography using spherical mds. *BMC medical imaging*, 10(1):12, 2010.

[GLH13] C. Giannella, K. Liu, and Kargupta H. Breaching euclidean distance preserving data perturbation using few known inputs. *Data and Knowledge Engineering*, 2013.

[HHC+14] Shu-Yu Hsu, Yingchieh Ho, Po-Yao Chang, Chauchin Su, and Chen-Yi Lee. A 48.6-to-105.2 μw machine learning assisted cardiac sensor soc for mobile healthcare applications. *IEEE Journal of Solid-State Circuits*, 49(4):801–811, 2014.

[HXY+16] Jian-Hua Huang, Hua-Lin Xie, Jun Yan, Dong-Sheng Cao, Hong-Mei Lu, Qing-Song Xu, and Yi-Zeng Liang. Correction: Interpretation of type 2 diabetes mellitus relevant gc-ms metabolomics fingerprints by using random forests. *Analytical Methods*, 8(8):1950–1951, 2016.

[JDXC+14] Lihong Jiang, Li Da Xu, Hongming Cai, Zuhai Jiang, Fenglin Bu, and Boyi Xu. An iot-oriented data storage framework in cloud computing platform. *IEEE Transactions on Industrial Informatics*, 10(2):1443–1451, 2014.

[JL84] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[JNY07] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. *ACM*, 2007.

[Jol17] Arnaud Joly. Exploiting random projections and sparsity with random forests and gradient boosting methods– application to multi-label and multi-output learning, random forest model compression and leveraging input sparsity. *arXiv preprint arXiv:1704.08067*, 2017.

[KEDR12] Adnan Mujahid Khan, Hesham El-Daly, and Nasir Rajpoot. Ranpec: Random projections with ensemble clustering for segmentation of tumor areas in breast histology images. In *Medical Image Understanding and Analysis*, pages 17–23, 2012.

[KKMM12a] K. Kenthapadi, A. Korolova, I. Mironov, and N. Mishra. Privacy via the Johnson-Lindenstrauss Transform. *ArXiv e-prints*, April 2012.

[KKMM12b] Krishnaram Kenthapadi, Aleksandra Korolova, Ilya Mironov, and Nina Mishra. Privacy via the johnson-lindenstrauss transform. *arXiv preprint arXiv:1204.2606*, 2012.

[KLR06] H. Kargupta, K. Liu, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge & Data Engineering*, 18:92–106, 01 2006.

[LF12] Li Liu and Paul Fieguth. Texture classification from random features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):574–586, 2012.

[LGK06] Kun Liu, Chris Giannella, and Hillol Kargupta. An attacker's view of distance preserving maps for privacy preserving data mining. In *Proceedings of the 10th European Conference on Principle and Practice of Knowledge Discovery in Databases*, PKDD'06, pages 297–308, Berlin, Heidelberg, 2006. Springer-Verlag.

[Liu07a] Kun Liu. *Multiplicative Data Perturbation for Privacy Preserving Data Mining*. PhD thesis, University of Maryland, 2007.

[Liu07b] Kun Liu. *Multiplicative Data Perturbation for Privacy Preserving Data Mining*. PhD thesis, University of Maryland, 2007.

[LJJ+09] Ronan A Lyons, Kerina H Jones, Gareth John, Caroline J Brooks, Jean-Philippe Verplancke, David V Ford, Ginevra Brown, and Ken Leake. The sail databank: linking multiple health and social care datasets. *BMC medical informatics and decision making*, 9(1):3, 2009.

[LKR06] Kun Liu, Hillol Kargupta, and Jessica Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering*, 18(1):92–106, 2006.

[LL15] In Lee and Kyoochun Lee. The internet of things (iot): Applications, investments, and challenges for enterprises. *Business Horizons*, 58(4):431–440, 2015.

[MP99] George D Magoulas and Andriana Prentza. Machine learning in medical applications. In *Advanced Course on Artificial Intelligence*, pages 300–307. Springer, 1999.

[MSDPC12] Daniele Miorandi, Sabrina Sicari, Francesco De Pellegrini, and Imrich Chlamtac. Internet of things: Vision, applications and research challenges. *Ad hoc networks*, 10(7):1497–1516, 2012.

[New09] Natalie Privacy Law Blog Newman. Netflix sued for largest voluntary privacy breach to date, 2009.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[Pal05] Mahesh Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005.

[RvdM14] J Rivera and R van der Meulen. Gartner says the internet of things will transform the data center. *Retrieved August*, 5:2014, 2014.

[SDG+14] Beata Strack, Jhonathan P. Deshazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014.

[SS98] Pierangela Samarati and Latanya Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS '98, pages 188–, New York, NY, USA, 1998. ACM.

[Tan16] Colin Tankard. What the gdpr means for businesses. *Network Security*, 2016(6):5–8, 2016.

[Wik14] Suanu Bliss Wikina. What caused the breach? an examination of use of information technology and health data breaches. *Perspectives in health information management*, 11(Fall), 2014.

[XLZW16] Haozhe Xie, Jie Li, Qiaosheng Zhang, and Yadong Wang. Comparison among dimensionality reduction techniques based on random projection for cancer classification. *Computational biology and chemistry*, 65:165–172, 2016.

[ZWC+09] Evangelia I Zacharaki, Sumei Wang, Sanjeev Chawla, Dong Soo Yoo, Ronald Wolf, Elias R Melhem, and Christos Davatzikos. Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme. *Magnetic resonance in medicine*, 62(6):1609–1618, 2009.