Data & Knowledge Engineering 116 (2018) 21–41

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Privacy-preserving collaborative fuzzy clustering

Lingjuan Lyu[a,*], James C. Bezdek[b], Yee Wei Law[c], Xuanli He[b], Marimuthu Palaniswami[a]

[a] Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, Australia
[b] School of Computing and Information Systems, The University of Melbourne, Parkville, Australia
[c] School of Engineering, University of South Australia, Mawson Lakes, Australia

ARTICLE INFO

Keywords:
Participatory sensing
Collaborative learning
Clustering
Privacy-preserving
Randomisation

ABSTRACT

The proliferation of Internet of Things devices has contributed to the emergence of *participatory sensing* (PS), where multiple individuals collect and report their data to a third-party data mining cloud service for analysis. The need for the participants to collaborate with each other for this analysis gives rise to the concept of *collaborative learning*. However, the possibility of the cloud service being semi-honest poses a key challenge: preserving the participants' privacy.

In this paper, we address this challenge with a two-stage scheme called RG + RP: in the first stage, each participant perturbs his/her data by passing the data through a nonlinear function called *repeated Gompertz* (RG); in the second stage, he/she then projects his/her perturbed data to a lower dimension in an (almost) distance-preserving manner, using a specific *random projection* (RP) matrix. The nonlinear RG function is designed to mitigate *maximum a posteriori* (MAP) estimation attacks, while random projection resists *independent component analysis* (ICA) attacks and ensures clustering accuracy. The proposed two-stage randomisation scheme is assessed in terms of its recovery resistance to MAP estimation attacks. Preliminary theoretical analysis as well as experimental results on synthetic and real-world datasets indicate that RG + RP has better recovery resistance to MAP estimation attacks than most state-of-the-art techniques. For clustering, *fuzzy c-means* (FCM) is used. Results using seven cluster validity indices, *root mean squared error* (RMSE) and accuracy ratio show that clustering results based on two-stage-perturbed data are comparable to the clustering results based on raw data — this confirms the utility of our privacy-preserving scheme when used with either FCM or HCM.

1. Introduction

The ubiquity of mobile sensing devices gave birth to *participatory sensing* (PS), a data crowdsourcing paradigm where participants “use evermore capable mobile phones and cloud services to collect and analyse systematic data for use in discovery” [1]. A closely related trend is *collaborative learning*, a data crowdsourcing paradigm where the participants not only “contribute individually collected training samples”, but also “collaboratively construct statistical models for tasks in pattern recognition” [2]. Essentially, PS is participant-oriented, sensing-focused and cloud-assisted collaborative learning.

In this work, we are concerned with the PS scenario where the participants are data owners that contribute their data, but rely on a third-party data mining cloud service to perform clustering-based analysis on their joint data (see Fig. 1). For the clustering operation, we consider specifically *fuzzy c-means* [3], because it is widely used and well established. In this scenario, the participants'

[*] Corresponding author.
E-mail address: llv@student.unimelb.edu.au (L. Lyu).

https://doi.org/10.1016/j.datak.2018.05.002
Received 17 November 2016; Received in revised form 18 November 2017; Accepted 5 May 2018
Available online 12 May 2018

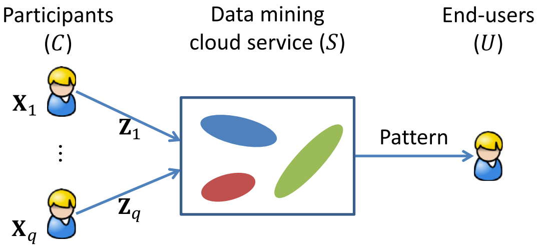


Fig. 1. The participatory sensing (PS) architecture, on which the research presented in this paper is based.

privacy becomes an issue because the cloud service might be *semi-honest*, i.e., although the cloud service does not deviate arbitrarily from the protocol, it might try to gain private information about the participants from their data. For participants that intend to contribute data, the privacy risks pose a major hurdle. Thus, our goal is to achieve *privacy-preserving collaborative fuzzy clustering*. The collaborative nature of the process stems from the common preprocessing step that each participant has to perform to preserve his/her privacy. Note that collaboration does not imply interaction.

There exist cryptographic techniques that allow data mining algorithms to be modified to process encrypted data; these techniques fall under the umbrella of *secure multiparty computation* (SMC) [4]. SMC achieves a high level of privacy and accuracy, but suffers from high computational complexity. Moreover, it requires synchronous interaction and several rounds of communications among participants, and hence suffer a lack of scalability. A more practical alternative to SMC is randomisation/perturbation, where each participant releases a perturbed version of his/her private data to the cloud service, with guarantees that the original sensitive information cannot be reidentified/recovered while the analytic properties of the data are maintained. Randomisation-based approaches aim to trade off privacy and accuracy for scalability, by perturbing data in a computationally efficient way that prevents the original data from being recovered accurately, but does not significantly affect the accuracy of data mining. Understandably, the availability of the perturbed data to the cloud service in plaintext (after decryption) and the possibility of the cloud service colluding with any subset of the participants mean randomised data are subject to statistical data recovery attacks [5–8]. The challenge addressed here is designing a randomisation-based PS scheme for fuzzy clustering that is resistant to data recovery by colluding attackers.

For privacy preservation, we explore the possibility of combining nonlinear transformation and multiplicative random projection matrices to construct a new representation of the data. A two-stage scheme called RG + RP is proposed, with the aim of achieving trade-off between privacy and accuracy: given the perturbed data, the server cannot infer the original values, even under collusion. However, this must be done without over-sacrificing accuracy, i.e., the results based on the perturbed data should be close to the corresponding results using the original data. To fulfill these requirements, in the first stage, each participant perturbs his/her data by passing the data through a nonlinear function called *repeated Gompertz* (RG) [9]; and in the second stage, each participant's data is projected to a lower dimension using a *random projection* (RP) matrix [10]. The nonlinear function is designed to condition the probability density function (pdf) of the perturbed data to thwart *maximum a posteriori* (MAP) estimation attacks [7], whereas the random projection operation compresses the data in a manner that on average preserves Euclidean distances between pairs of data points as per the Johnson-Lindenstrauss Lemma [10]. Furthermore, random projection is known to be resistant to *independent component analysis* (ICA) attacks [10].

For evaluating privacy preservation, we use the *ε-recovery rate* metric [9], which measures how much data can be recovered within a relative error of ε (see Section 5).

For evaluating distortion after perturbation, we calculate the inner products and Euclidean distances between raw data points, and compare them to the inner products and Euclidean distances between perturbed data points, in terms of the *root mean squared error* (RMSE). Accuracy ratio between clustering accuracy based on the perturbed data and raw data is evaluated, which reflects the impact of perturbation on the raw data. For pattern recognition, seven *clustering validity indices* (CVIs) are used [11]: (i) Adjusted Rand index, (ii) Rand index, (iii) mutual information, (iv) normalised mutual information, (v) variation of information, (vi) normalised variation of information, and (vii) Jaccard index.

Our contributions can be summarised as follows. First, we propose a two-stage perturbation scheme that can demonstrably preserve the privacy of both normal and anomalous data records under MAP estimation attacks. Secondly, we show that a uniform random transformation fails to preserve Euclidean distances between data points, and consequently, we rely on the stochastic distance-preserving property of random projection to maintain clustering accuracy. Thirdly, we present a framework for validating post-perturbation clustering results in terms of RMSE, clustering accuracy and CVIs. Finally, we present experimental results that show RG + RP is applicable to large and high-dimensional datasets.

The rest of this article is organised as follows. Section 2 discusses related work. Section 3 discusses multiplicative perturbation. Section 4 presents the details of our privacy-preserving scheme, RG + RP. Section 5 analyses the privacy-preserving properties of RG + RP. Section 6 analyses the post-perturbation clustering accuracy of fuzzy *c*-means. Section 7 concludes. For the readers' convenience, Tables 1 and 2 contain a list of acronyms and a list of symbols used in this article.

Table 1
Table of acronyms.

Acronym	Meaning
CI	Consensus index
CVI	Clustering validity indices
FCM	Fuzzy *c*-means
HCM	Hard *c*-means
ICA	Independent component analysis
i.i.d.	independent and identically distributed
MAP	Maximum a posteriori
MI	Mutual information
pdf	Probability density function
PS	Participatory sensing
RG	repeated Gompertz
RMSE	Root mean squared error
RP	Random projection
RT	Random transformation

Table 2
Table of symbols.

Symbol	Meaning
n	Number of attributes
m	Number of samples
$\mathscr{N}$	Nonlinear function
E	Expectation operator
Var	Variance operator
$N(\mu, \sigma^2)$	Normal/Gaussian distribution with mean μ and variance σ^2
$U(a, b)$	Uniform distribution on the interval $[a, b]$
$\mathbf{N}$	Contingency table (see Section 6.2)

2. Related work

This section discusses related work first in randomisation-based privacy-preserving schemes, then in collaborative clustering.

2.1. Randomisation-based privacy-preserving schemes

In general, semantic privacy criteria 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 exponential distributed noise to the query results [12]. 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, which necessitates the help of cryptographic mechanisms as evidenced by the schemes below:

- Dwork et al.'s scheme [13] requires participants to communicate with each other, and use verifiable secret sharing to collaboratively generate shares of random noise, imposing computational and communication costs on the order of 2^t, where t is the estimated maximum number of dishonest participants [14].
- Shi et al.'s scheme [15] enables participants to upload encrypted values to a data aggregator, who computes the sum of the encrypted values. These values are perturbed with geometric noise (which is approximately the discrete version of Laplace noise) to satisfy (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 [16] 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. Our scheme uses the criterion *recovery resistance*, which is defined in Section 5. This criterion is based on the *recovery rate* metric used in Sang et al.'s innovative study of attacks on randomisation-based schemes [7].

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

Additive perturbation: Agrawal and Srikant [17] proposed perturbing data matrix $\mathbf{X}$ by adding independent and identically distributed (i.i.d.) noise $\mathbf{R}$ as $\mathbf{X} + \mathbf{R}$, but Huang et al. [5] questioned the use of additive noise and pointed out it can be filtered out using spectral filtering techniques, leading to privacy breach. The general principle is to correlate the additive noise with the original data [18], but a participant can infer the data of another participant if their data happens to be correlated [19].

Multiplicative perturbation: The original data matrix $\mathbf{X}$ is perturbed by multiplying it with some random noise matrix $\mathbf{R}$ as $\mathbf{RX}$, and only the perturbed version is released for data analysis. Well-known multiplicative schemes include:

- *Rotation perturbation*: The noise matrix $\mathbf{R}$ is an orthogonal matrix whose columns and rows are orthogonal unit vectors [20]. This scheme is vulnerable to "known-input attacks" [8], where an attacker can recover the original data from its perturbed version with just a few leaked inputs.
- *Random projection* (RP): The idea of random projection originated in the following seminal theorem:

Lemma 1. (Johnson-Lindenstrauss Lemma [21]) For any set M of $m = |M|$ data points in $\mathbb{R}^n$, given $\lambda > 0$ and $w = \Omega\left(\frac{\ln m}{\lambda^2}\right)$ ($w \geq \frac{4\ln m}{\lambda^2/2-\lambda^3/3}$), there exists a map that embeds the set into w -dimensional subspace $\mathbb{R}^w$, such that the pairwise distance of any two points is maintained within an arbitrarily small factor 1 ± λ, i.e., for all $\vec{x_1}, \vec{x_2} \in \mathbb{R}^n$, a linear projection $\mathbf{P} \in \mathbb{R}^{w\times n}$ satisfies:

$$(1-\lambda)\|\vec{x_1}-\vec{x_2}\|_2^2 \leq \|\mathbf{P}\vec{x_1}-\mathbf{P}\vec{x_2}\|_2^2 \leq (1+\lambda)\|\vec{x_1}-\vec{x_2}\|_2^2. \tag{1}$$

Lemma 1 is an existence theorem that basically says given a matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$, where n is the number of features and m is the number of data records, there exists a matrix $\mathbf{T} \in \mathbb{R}^{w\times n}$ such that $\mathbf{Y} = \mathbf{TX} \in \mathbb{R}^{w\times m}$ has similar inter-column distances to $\mathbf{X}$'s, provided $w \geq \frac{4\ln m}{\lambda^2/2-\lambda^3/3}$. RP has been used since the 1990s for dimensionality reduction [22]. This property makes RP an excellent perturbation method to be used in conjunction with distance-based data mining techniques. RP matrices can be generated in multiple ways, including:

– Drawing each entry of the matrix independently from the normal distribution N(0, $1/w$) [23,24].
– Drawing each entry of the matrix independently and uniformly at random from $\left\{-\frac{1}{\sqrt{w}}, \frac{1}{\sqrt{w}}\right\}$ [25].
– Drawing each entry of the matrix independently from $\left\{-\sqrt{\frac{3}{w}}, 0, \sqrt{\frac{3}{w}}\right\}$ at probabilities $\left\{\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right\}$ respectively [25].
– Generating each entry of the matrix by randomised hashing, with the resultant matrix being a sparse matrix [26,27].

Liu et al. [10] showed that if the original data set with n attributes is multiplied by a $w \times n$ ($w < n$) mixing matrix, which is random and orthogonal in expectation, then the perturbed data with reduced dimension can be released for data mining while preserving much of its distance-related characteristics. However, Sang et al. [7] and Liu et al. [28] addressed the possible risks of RP by studying how well an attacker can recover the original data from the transformed data with the aid of prior information: if the original data follows a multivariate Gaussian distribution, a large portion of the data can be reconstructed via *maximum a posteriori* (MAP) estimation. In order to resist MAP estimation attacks, we introduce nonlinear transformation as a pre-RP step to condition the pdf of the perturbed data [29].

- *Uniform random transformation* (RT): The noise matrix is defined as a matrix whose elements are independently sampled from the uniform distribution U(0, 1) [30]. Unlike rotation perturbation and RP, RT does not aim for orthogonality or distance preservation.

Geometric perturbation: A mix of additive and multiplicative perturbations are used in geometric perturbation, where the data matrix $\mathbf{X}$ is mapped to $\mathbf{RX} + \mathbf{\Phi} + \mathbf{\Delta}$, where $\mathbf{R}$ refers to 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 [20]. It is known that without $\mathbf{\Delta}$, geometric perturbation is vulnerable to "known input attacks" [8], but there are no general results on how the $\mathbf{\Delta}$ term reduces the effectiveness of these attacks.

Nonlinear transformation: It is clear from Ref. [7] that randomisation must have nonlinear components, i.e., on top of linear mapping, nonlinear transformation is required. First proposed by Bhaduri et al. [31], randomisation in the most general form takes the form $\mathbf{B} + \mathbf{Q}\cdot\mathcal{N}(\mathbf{A} + \mathbf{RX})$, where $\mathbf{B}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{R}$ are random matrices, and $\mathcal{N}$ is a bounded nonlinear function. The tanh function is found to preserve the distance between normal data points, but collapses the distance between outliers, making the function suitable for privacy-preserving pattern learning, such as anomaly detection, provided only the privacy of anomalous records needs to be protected.

Lyu et al. [9] proposed an improved two-stage perturbation scheme which relies on a nonlinear transformation and participant-specific U(0, 1)-distributed multiplication matrix to resist both MAP and ICA attacks. The innovations include (i) the use of participant-specific perturbation matrices, and (ii) the use of the *repeated Gompertz* (RG) function as the nonlinear transformation function with many-to-one mapping for a subset of the domain — a property effective for protecting both anomalous and normal records from MAP estimation attacks. However, a participant-specific perturbation matrix does not preserve the distance between data points, limiting its application to distance-based clustering tasks.

Nevertheless, none of the methods above achieve a good trade-off between privacy and accuracy for distance-based clustering. In this paper, we use the RG function as the nonlinear transformation function in the first stage to resist MAP estimation attacks, and rely on RP in the second stage to resist ICA attacks and ensure clustering accuracy. This paper shows how to successfully tailor the nonlinear transformation and random projection to privacy-preserving collaborative clustering applications.

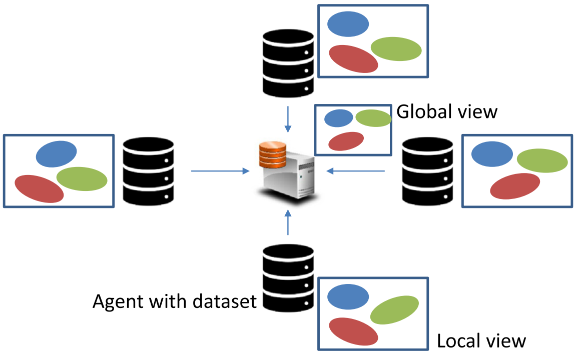


Fig. 2. The centralised collaborative clustering architecture. See Section 2.2 for related work based on this architecture.

2.2. *Collaborative clustering*

Historically, the term “collaborative clustering” has been used to refer to multiple things [32]. The term has been used to refer to a way of combining clustering methods [33] that improves upon ensemble clustering algorithms. This meaning of “collaborative clustering” is irrelevant to the work presented here.

The term “collaborative clustering” has also been used to refer to multiparty clustering based on either the centralised architecture in Fig. 2 or the distributed architecture in Fig. 3. In this context, collaborative clustering is used when multiple parties want to refine their local clustering model or obtain a global clustering model without exchanging data. In this context, collaborative clustering is also called multiple-view clustering [34]. The collaborative clustering paradigm addressed in this article can be viewed as a PS-friendly extension of the centralised collaborative clustering paradigm to the specific scenario where the agents have as few as one record of the dataset, and the luxury of being able to offload intensive computation to the cloud. The distinctions among the different variants of collaborative clustering are summarised in Table 3.

There are distributed algorithms, as shown in Table 3, for improving local clustering through collaboration specified by so-called interaction coefficients or interaction levels [37,38]. In these schemes, each agent has a local view of how its data points are clustered, rather than a global view of how all data points are clustered. In other words, these schemes do not exchange data, but they do not generate any global model either.

There are algorithms for generating a global model, but some only work on generative models [36], some have unknown communication complexity [34], and some require pairwise communications and have nonnegligible communication complexity [38]. For example, the collaborative fuzzy clustering scheme based on so-called parallel fuzzy c-means [40] has a communication complexity of $O(c_{\max} \cdot d \cdot t \cdot P)$, where $c_{\max}$ is the maximum number of clusters, d is the number of attributes, t is the number of collaboration rounds, and P is the number of agents [38]. Yet some are tailored to peer-to-peer networks [39], which do not reflect the participatory sensing scenario.

In stark contrast, there is hardly any work done on privacy-preserving PS-friendly collaborative clustering, which is a relatively new collaborative clustering paradigm.

3. Comparing multiplicative perturbation techniques

Euclidean distance-preserving data perturbation [20,41] has been getting attention because it facilitates privacy/accuracy trade-

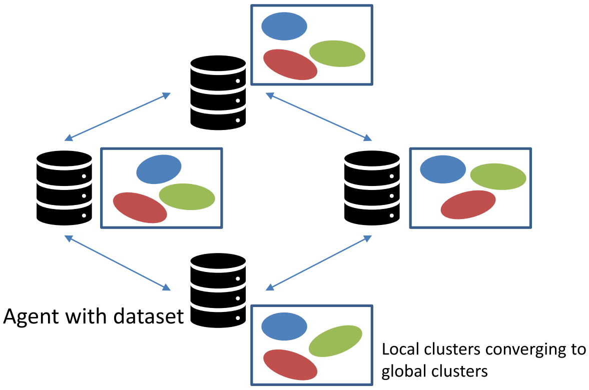


Fig. 3. The distributed collaborative clustering architecture. See Section 2.2 for related work based on this architecture.

Table 3
Comparing different variants of collaborative clustering.

Collaborative clustering	PS-friendly	Centralised	Distributed
Architecture	Centralised as in Fig. 1	Centralised as in Fig. 2	Distributed as in Fig. 3
Global model	Yes	Yes	Depends
Sample references	Non-privacy-preserving: [35]:	Privacy-preserving: [34,36]	Privacy-preserving: [37–39]:

off. Euclidean distance-preserving data perturbation can be illustrated as follows. An organization has a private, real-valued dataset $\mathbf{X}$ (row: feature, column: data record) and wishes to make it publicly available for data analysis while keeping the individual records private. To accomplish this, $\mathbf{Y} = \mathbf{TX}$ is released to the public, where $\mathbf{T}$ preserves Euclidean distances between columns and is only known to the data owner. With this nice property, many useful data mining algorithms, with only a minor modification, can be applied to $\mathbf{Y}$ and produce nearly the same patterns that can be extracted from $\mathbf{X}$.

Random projection is known to preserve Euclidean distance in expectation. For a more rigorous analysis, Let $\mathbf{T} = [t_{ij}]$, where $i = 1,\ldots,w, j = 1,\ldots,n$, and $w < n$. Given some distribution D, if $t_{ij} \sim$ D is i.i.d., we write $\mathbf{T} \sim \mathrm{D}_{w\times n}$. In Proposition 1, we consider two types of distributions (and hence two types of $\mathbf{T}$), in terms of their ability to preserve Euclidean distances and inner products.

Proposition 1. *Suppose $\boldsymbol{T}$ is a $w \times n$ multiplicative perturbation matrix, where $w < n$. Then,*

Case 1. *If $\mathbf{T} \sim \mathsf{N}_{w\times n}(0, \sigma_t^2)$, where $\mathsf{N}_{w\times n}(0, \sigma_t^2)$ is the zero-mean Gaussian distribution with variance $\sigma_t^2 = \frac{1}{w}$, then both the inner products and Euclidean distances are preserved in expectation, i.e., the error of the inner product produced by random projection is zero on average, and the variance is inversely proportional to the reduced dimensionality w.*

Case 2. *If $\boldsymbol{T} \sim U_{w\times n}(0, 1)$, where $U(0, 1)$ is the uniform distribution on the interval $[0, 1]$, then both the Euclidean distances and inner products between data points are not preserved.*

Proof. Let t_{ij} and ε_{ij} *be* the (i, j)-th entry of projection matrix $\mathbf{T}$ and $\mathbf{T}^\top\mathbf{T}$ respectively. We consider the two cases in turn:

Case 1. $t_{ij} \sim \mathsf{N}(0, \sigma_t^2) \Longrightarrow \mathsf{E}[t_{ij}^2] = \sigma_t^2 = \frac{1}{w}$. It follows that $\mathsf{E}[\varepsilon_{ii}] = w\sigma_t^2 = 1$, $\mathsf{Var}[\varepsilon_{ii}] = 2w\sigma_t^4 = \frac{2}{w}, \forall i$; and $\mathsf{E}[\varepsilon_{ij}] = 0$, $\mathsf{Var}[\varepsilon_{ij}] = w\sigma_t^4 = \frac{1}{w}$, $\forall i \neq j$. Therefore, $E[\boldsymbol{T}^\top\boldsymbol{T}] = \mathbf{I}$.

Let $\vec{x}, \vec{y} \in \mathbb{R}^n$. Suppose $\vec{x}$ and $\vec{y}$ are projected to $\vec{u} = \mathbf{T}\vec{x}$ and $\vec{v} = \mathbf{T}\vec{y}$ respectively. Then,

$$\begin{aligned} \mathsf{E}[\vec{u}^{\top}\vec{v} - \vec{x}^{\top}\vec{y}] &= \mathsf{E}[\vec{x}^{\top}\mathbf{T}^{\top}\mathbf{T}\vec{y} - \vec{x}^{\top}\vec{y}] = \vec{x}^{\top}\mathsf{E}[\mathbf{T}^{\top}\mathbf{T}]\vec{y} - \vec{x}^{\top}\vec{y} = 0, \\ \mathsf{Var}[\vec{u}^{\top}\vec{v} - \vec{x}^{\top}\vec{y}] &= \tfrac{1}{w}\{\textstyle\sum_i x_i^2 \sum_i y_i^2 + (\sum_i x_i y_i)^2\}. \end{aligned}$$

In particular, if both $\vec{x}$ and $\vec{y}$ are normalised, then $(\sum_i x_i^2)(\sum_i y_i^2) = 1$, $(\sum_i x_i y_i)^2 \leq 1$, and the upper bound of the variance becomes:

$$\mathsf{Var}[\vec{u}^{\top}\vec{v} - \vec{x}^{\top}\vec{y}] \leq \frac{2}{w},$$

which is inversely proportional to the reduced dimension w. Applying the results above to vectors $\vec{u} - \vec{v}$ and $\vec{x} - \vec{y}$, we have

$$\mathsf{E}[(\vec{u} - \vec{v})^{\top}(\vec{u} - \vec{v}) - (\vec{x} - \vec{y})^{\top}(\vec{x} - \vec{y})] = 0 \Longrightarrow \mathsf{E}[\|\vec{u} - \vec{v}\|_2^2 - \|\vec{x} - \vec{y}\|_2^2] = 0.$$

If both $\vec{x}$ and $\vec{y}$ are normalised, then $\mathsf{Var}[\|\vec{u} - \vec{v}\|_2^2 - \|\vec{x} - \vec{y}\|_2^2] \leq \frac{32}{w}$. The detailed proofs are provided in [42, Appendix 5.6.1].

Case 2. $t_{ij} \sim \mathsf{U}(0, 1) \Longrightarrow \mathsf{E}[t_{ij}] = \frac{1}{2}$, and $\mathsf{E}[t_{ij}^2] = \frac{1}{3}$. As $\varepsilon_{ii} = \sum_{k=1}^{w} t_{ki}^2$ and $\varepsilon_{ij} = \sum_{k=1}^{w} t_{ki}t_{kj}$, $\forall\ i \neq j$, thus $\mathsf{E}[\varepsilon_{ii}] = \mathsf{E}[\sum_{k=1}^{w} t_{ki}^2] = w\mathsf{E}[t_{ki}^2] = \frac{w}{3}$, and $\mathsf{E}[\varepsilon_{ij}] = \mathsf{E}[\sum_{k=1}^{w} t_{ki}t_{kj}] = \sum_{k=1}^{w} \mathsf{E}[t_{ki}]\mathsf{E}[t_{kj}] = \frac{w}{4}$. Therefore,

$$\mathsf{E}[\mathbf{T}^{\top}\mathbf{T}] = [\tau_{ij}], \qquad \text{where } \tau_{ij} = \begin{cases} \frac{w}{4}, & \text{if } i \neq j; \\ \frac{w}{3}, & \text{if } i = j. \end{cases}$$

It then follows that

$$\mathsf{E}[\vec{u}^{\top}\vec{v} - \vec{x}^{\top}\vec{y}] = \mathsf{E}[\vec{x}^{\top}\mathbf{T}^{\top}\mathbf{T}\vec{y} - \vec{x}^{\top}\vec{y}] = \vec{x}^{\top}\mathsf{E}[\mathbf{T}^{\top}\mathbf{T}]\vec{y} - \vec{x}^{\top}\vec{y} = \vec{x}^{\top}\vec{y}\,(E\,[\mathbf{T}^{\top}\mathbf{T}] - 1) \neq 0.$$

Hence, both the Euclidean distances and inner products between data points are not preserved.□

Proposition 1 shows that a normally distributed RP matrix preserves both Euclidean distances and inner products in expectation, hence it is beneficial for distance-based analytics such as clustering.

4. The RG + RP scheme

As depicted in Fig. 1, the general participatory sensing architecture comprises a set of participants $\mathscr{C} = \{c_i | i = 1,\ldots,q\}$, a data mining cloud service $\mathscr{S}$, and a set of end-users $\mathscr{U}$. In our privacy model, the cloud service is assumed to be semi-honest, 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. Based on the state of the art in privacy-preserving data mining, the following design criteria are considered:

- **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 [7]. A nonlinear transformation can be applied to prevent this reduction by conditioning the pdf.
- **Resilience to ICA attacks**: Independent Component Analysis (ICA) aims to discover independent hidden factors that underlie a set of linear or nonlinear mixtures of some unknown variables. ICA attacks normally recover the original signals from only the observed mixture by a filter. To counter the possible prerequisites [42] for an ICA attack to succeed, we rely on random projection-based perturbation to enforce the reduced dimension to be lower than the half of the original dimension.

4.1. RG+RP

All participants generate based on a pre-distributed shared seed, a $w \times n$ random projection (RP) matrix $\mathbf{T} \sim \mathrm{N}_{w\times n}(0, 1/w)$, where $w < n$. Suppose participant c_i with n features and m_i records is contributing data $\mathbf{X}_i \in \mathbb{R}^{n\times m_i}$ to the cloud service S for clustering. 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 $\mathscr{N}$ element-wise:

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

$\mathscr{N}$ is chosen to be the repeated Gompertz(RG) function:

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

where the parameters a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2 are defined in Fig. 4, and $u()$ is the Heaviside step function. The derivation of the function parameters is explained in Section 5. Fig. 4 plots different nonlinear perturbation functions for comparison.

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

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

The participant then sends $\mathbf{Z}_i$ to the cloud service $\mathscr{S}$. Once $\mathscr{S}$ 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 conducts clustering on $\mathbf{Z}_{\text{all}}$. The clustering results provide a general view of data patterns of the combined participants' records. The pseudocode for the participatory clustering procedure is shown in Algorithm 1. The role of $\mathscr{S}$ is to conduct clustering on the received $\mathbf{Z}_{\text{all}}$. End-users can get access to the pattern results for analysis. Our perturbation scheme is independent of the clustering algorithm used, fuzzy clustering 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

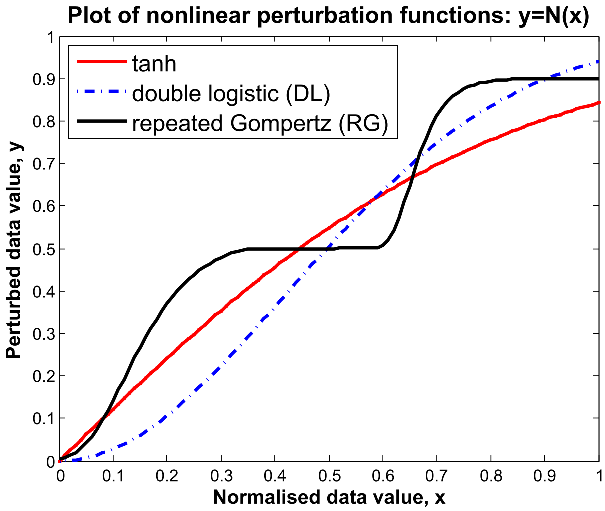


Fig. 4. Parameters of the repeated Gompertz function, and a plot of different nonlinear perturbation functions. The tanh function is $\mathscr{N}(x) = \tanh(\beta_t x)$, where $\beta_t \approx 1.23$ [29]. The double logistic function is $\mathscr{N}(x) = 1 - \exp(-\beta_{dl} x^2)$, where $\beta_{dl} \approx 2.81$ [29]. The repeated Gompertz function is defined in Eq. (3).

Algorithm 1
The participatory clustering procedure.

```
Role: Participant
Require: $\mathscr{N}$
    Generate RP matrix $\mathbf{T} \sim \mathrm{N}_{w \times n}(0, 1/w)$, based on a pre-distributed shared seed
    $\mathbf{Z}_i \leftarrow \mathbf{T}(\mathscr{N}(\mathbf{X}_i))$
    Send $\mathbf{Z}_i$ to the cloud service
Role: Cloud service
    Receives $\mathbf{Z}_i$, $i = 1,\ldots,q$
    $\mathbf{Z}_{\text{all}} \leftarrow [\mathbf{Z}_1|\mathbf{Z}_2|\cdots|\mathbf{Z}_q]$
    Conduct clustering on $\mathbf{Z}_{\text{all}}$
    Send clustering results to the end-users
Role: End-user
    Receive clustering results from cloud service
```

Below, we use a two-participant scenario as an example. Let $\mathbf{X}_1$ be an $n \times m_1$ data matrix owned by Alice, and $\mathbf{X}_2$ be an $n \times m_2$ matrix owned by Bob. A third-party cloud service is available for conducting fuzzy clustering on the union of these two datasets, $[\mathbf{X}_1|\mathbf{X}_2]$, without directly accessing the original data. The procedure is detailed below:

1. Alice and Bob generate $\mathbf{T} \sim \mathrm{N}_{w \times n}(0, 1/w)$.
2. Alice and Bob transform $\mathbf{X}_1$ and $\mathbf{X}_2$ to

$$\mathbf{Y}_1 = \mathscr{N}(\mathbf{X}_1) \quad \text{and} \quad \mathbf{Y}_2 = \mathscr{N}(\mathbf{X}_2) \quad \text{respectively.}$$

3. Project $\mathbf{Y}_1$ and $\mathbf{Y}_2$ onto $\mathbb{R}^w$ using $\mathbf{T}$ and then release the perturbed versions

$$\mathbf{Z}_1 = \mathbf{T}\mathbf{Y}_1 \quad \text{and} \quad \mathbf{Z}_2 = \mathbf{T}\mathbf{Y}_2 \quad \text{respectively.}$$

4. The third-party cloud service performs fuzzy clustering over the concatenated data set $[\mathbf{Z}_1|\mathbf{Z}_2]$.

5. Privacy analysis

We consider the privacy-preserving properties of RG + RP in terms of its ability to resist *maximum a posteriori* (MAP) estimation attacks. MAP estimation is based on solid Bayesian statistics and is more general than maximum likelihood estimation because the former takes the prior distribution into account. It often produces estimation errors that are not much higher than the minimum mean square error; and it is relatively easy to derive the conditional probability density function in the multiplicative data perturbation scenario.

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 $\mathbf{T}\vec{y} = \vec{z}$ to work with, where $\vec{z}$ is a projection of vector $\vec{y}$. An underdetermined system cannot be solved for $\vec{y}$ exactly, but given sufficient prior information about $\vec{y}$, an approximation of the true $\vec{y}$ may be attainable. We characterise an attack by the extent of prior information available to the attacker.

In most realistic scenarios, the attacker has some additional prior knowledge which can potentially be used effectively for breaching privacy. We consider the following two scenarios of prior knowledge. 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 [8]. 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.

To measure the effectiveness of the reference attack, the recovery rate is defined as follows. If the recovered version for a data vector $\vec{x}$ is $\hat{\vec{x}}$, then the *relative error* is $\xi \overset{\text{def}}{=} \left\| \hat{\vec{x}} - \vec{x} \right\|_2 / \|\vec{x}\|_2$, where $\|\cdot\|_2$ is the Euclidean norm.

Denote the joint distribution of ξ and $\vec{x}$ by $p_{\Xi,\vec{X}}(\xi, \vec{x})$, then we define the *ε-recovery rate* with respect to the perturbation algorithm and attack as

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

where $D_{\vec{x}}$ is the domain of the data vector, and $\vec{x}$ is normalised. The joint distribution $p_{\Xi,\vec{X}}$ depends on the attack $\mathscr{A}$ 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 $\vec{x} \in \mathbb{R}^n$ as input and produces $\vec{z} \in \mathbb{R}^w$ as output is (ε, δ)-*recovery resistant* with respect to p_D and attack algorithm $\mathscr{A}$ if $r_\varepsilon(\mathscr{A}, p_D) = \delta$.

Suppose the attacker is targeting a particular participant by trying to solve $\mathbf{Z} = \mathbf{TY}$ for $\mathbf{Y}$. We consider two cases of prior knowledge about random matrix: where $\mathbf{T}$ is known, and where $\mathbf{T}$ is unknown.

5.1. Random matrix *T* is known

In the worst case, the random matrix $\mathbf{T}$ is disclosed exactly to the attacker, i.e., the specific realization of the random matrix is disclosed, is has been demonstrated in Ref. [10] that it is impossible for the adversary to find the exact value of any element in the original data, and an estimated one may be used instead. A more relaxed scenario is when the dimensionality and the distribution of the random matrix $\mathbf{T}$ are disclosed, the adversary cannot identify the original data by a random guess of the random matrix, all he can get is approximately a null matrix with all entries being around 0. 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 $\vec{z}$ represent a column of $\mathbf{Z}$, and $\vec{y}$ represent a column of $\mathbf{Y}$. The MAP estimate of $\vec{y}$, given $\mathbf{T}$ and $\vec{z}$, is

$$\hat{\vec{y}} = \arg\max_{\vec{y}} \; p(\vec{y}\,|\vec{z}, \mathbf{T}) = \arg\max_{\vec{y}} \; \frac{p(\vec{z} \mid \mathbf{T}, \vec{y})p(\mathbf{T})p(\vec{y})}{p(\vec{z} \mid \mathbf{T})p(\mathbf{T})} = \arg\max_{\vec{y} \in \mathscr{Y}} \; \frac{p(\vec{y})}{\int_{\mathbb{R}^n} p(\vec{z} \mid \mathbf{T}, \vec{y})\mathrm{d}\vec{y}} = \arg\max_{\vec{y} \in \mathscr{Y}} \; p(\vec{y}), \tag{6}$$

where $\mathscr{Y} = \{\vec{y} : \vec{z} = \mathbf{T}\vec{y}\}$. Note:

- The factor $p(\vec{z}|\mathbf{T}, \vec{y})$ translates to the constraint $\vec{y} \in \mathscr{Y}$.
- The integral in the denominator does not contribute towards maximising $\vec{y}$.

If $\vec{y}$ is n-variate Gaussian with a positive definite covariance matrix, then Eq. (6) becomes an easily solvable quadratic programming problem [7, Theorem 1]. The key is to design a nonlinear function $\mathscr{N}$ that transforms a potentially Gaussian data distribution to a distribution that deters accurate solution of Eq. (6).

In this paper, we use the repeated Gompertz(RG) function defined in Eq. (3) as the nonlinear function. The traditional Gompertz function takes the standard form:

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

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. The following explains how the proposed function is derived. As $\tanh(\beta_t x)$ is good for protecting anomalous data points, the slopes of repeated Gompertz function at $x = 0$ and $x = 1$ approximate those of $\tanh(\beta_t x)$. As the slope becomes steeper, corresponding to a larger absolute value of c, the less invertible the function, and therefore the higher degree of privacy, because it tends to produce a many-to-one mapping. In order to protect normal data points, the repeated Gompertz function is also designed to have a flat middle section so that for that section the function cannot be inverted. Through extensive experimentation, the geometry of the function in Fig. 4 have been empirically found to be effective for hampering Bayesian estimation attacks for 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. 4. This compositional structure inspired the name “repeated Gompertz”.

5.2. Random matrix *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

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

where $\Theta = \{(\mathbf{T}, \mathbf{Y}): \mathbf{Z} = \mathbf{TY}\}$. 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). When $\mathbf{T}$ is known, the repeated Gompertz function is designed to make data recovery via Eq. (6) difficult. 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.

5.3. Underdetermined independent component analysis (UICA)

ICA is a statistical technique which aims to represent a set of random variables as linear combinations of statistically independent

Table 4
Privacy experimental datasets.

Datasets	#records(m)	Upspace dimension	Downspace dimension
purely Gaussian	5000	15	8
purely Laplace	5000	15	8
Abalone	4177	8	4
HTRU2	17898	8	4
Adult	48842	123	62
HAR	7352	561	281
lymphoma	45	4026	2013

component variables. The aim of an ICA attack is to design a filter that can recover the original signals from only the observed mixture. ICA can separate out **T** and **Y**, knowing only their product **Z** = **TY**, provided (i) The number of observed attributes must be at least as large as the independent attributes, $w \geq n$; (ii) the attributes are independent; (iii) at most one of the attributes is Gaussian; (iv) **T** must have full column rank. To resist an ICA attack, we enforce $w < n$, namely projecting data to a lower-dimensional subspace to make the problem of ICA underdetermined/overcomplete. In this case, even if the perturbation matrix **T** is available, the independent components cannot be obtained. Moreover, as shown in Ref. [10], if $w \leq (n + 1)/2$, no linear filter can separate the observed mixture **Z**. It is demonstrated that an ICA attack cannot effectively breach the privacy of random projection-based perturbation, while the randomly generated projection matrix is likely to be more appropriate for protecting the privacy, compressing the data, and still maintaining its utility.

5.4. Privacy evaluation

This section presents the simulation and evaluation results of RG + RP in terms of its recovery resistance. Since it is resistant to ICA, the empirical experiments focus on the recovery rate of our two-stage scheme under the MAP estimation attack and compare it to earlier results [10,31] and [29]. Experiments are conducted on both synthetic and real datasets as shown in Table 4:

In order to evaluate the recovery resistance of RG + RP against the MAP estimation attacks, experimental results are provided 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:

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

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

To execute MAP estimation, the attacker can either apply the [7, 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 by using the leaked input samples and 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.

The four schemes shown in Table 5 are evaluated in the *worst-case* scenario where the attacker knows exactly the victim's perturbation matrix. The privacy results for different datasets are as follows.

Purely Gaussian datasets: As demonstrated in Fig. 5, RG + RP provides significantly higher recovery resistance for both normal and anomalous data compared to the other schemes, except for the 0.2-recovery rate for the anomalous data case, which is less effective than RP and tanh + RP. And it is slightly less effective than RP for the 0.1-recovery rate for the anomalous data case.

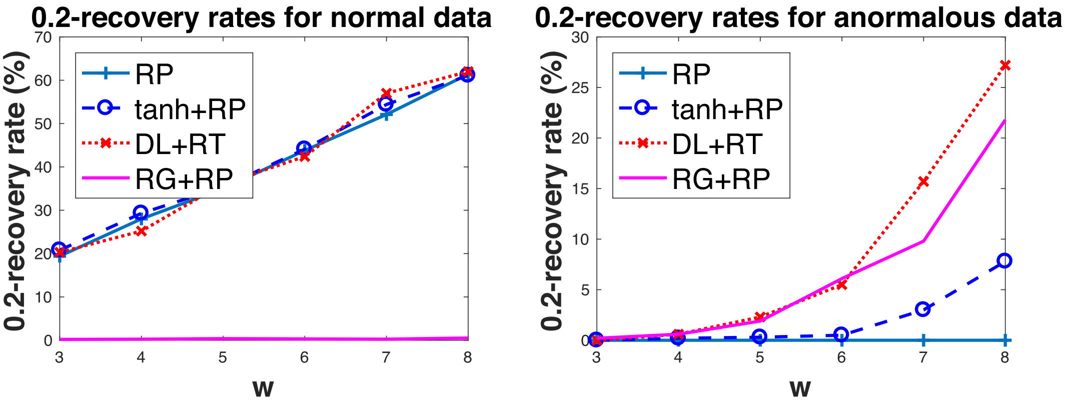
Purely Laplace datasets: Fig. 6 shows that RG + RP significantly outperforms other methods for Laplace datasets, and this is especially evident for normal data, except for the 0.2-recovery rate for the anomalous data case, which is slightly less effective than tanh + RP. Furthermore, the 0.1-recovery rate against RG + RP 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. 7, RG + RP also outperforms tanh + RP 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 + RP achieves (0.1, 0)-recovery resistance for the HAR and lymphoma datasets.

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

Table 5
Evaluated schemes.

Scheme	Nonlinear perturbation function (stage 1)	Linear projection matrix (stage 2)
RP [10]	none	$\mathbf{T} \sim \mathrm{N}_{w\times n}(0, 4)$
tanh + RP	tanh [31]	$\mathbf{T} \sim \mathrm{N}_{w\times n}(0, 1)$
DL + RT [29]	double logistic	$\mathbf{T} \sim \mathrm{U}_{w\times n}(0, 1)$
RG + RP	repeated Gompertz	$\mathbf{T} \sim \mathrm{N}_{w\times n}(0, 1/w)$



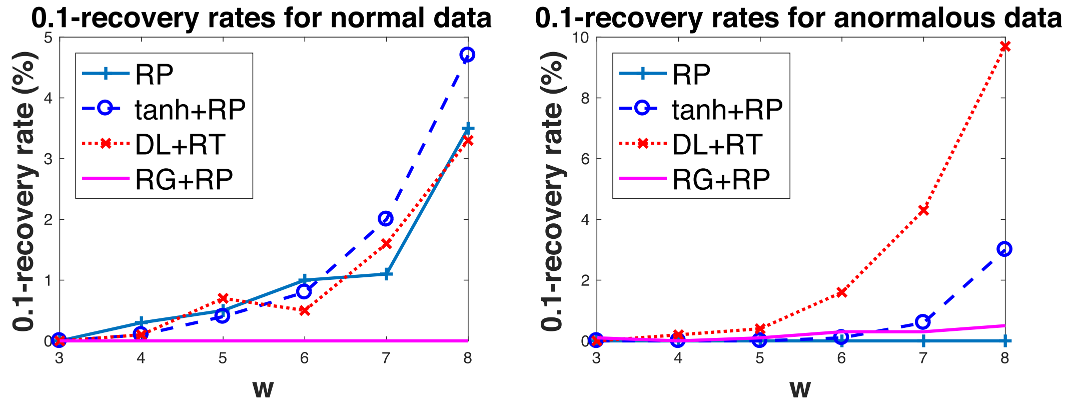


Fig. 5. Recovery rates of MAP estimation attacks against the evaluated schemes, on $w \times 1000$ data projected from 15×1000 normalised Gaussian-distributed data (zero mean, identity covariance matrix).

6. Fuzzy clustering and performance evaluation

6.1. Implementation and settings

In this paper, we rely on *fuzzy c-means* (FCM) [3] for clustering, which generates soft partitions for any set of numerical feature vector data. A fuzzy c-partition matrix $\mathbf{U}$ is shown in Fig. 8, where each entry $u_{ik} \in [0, 1]$ represents the membership of object (data point) o_k in cluster i. A partition matrix must meet two conditions:

Row sums: $\sum_k u_{ik} > 0$,
Column sums: $\sum_i u_{ik} = 1$.

We use the MATLAB implementation of FCM and specify values of c. More specifically, we run the FCM algorithm on each dataset to generate a set of soft partitions with the number of clusters ranging from $c_{true} - c_{true}/2$ to $c_{true} + c_{true}/2$. In order to reduce the influence of random initialization on the FCM algorithm, we generate 100 partitions for each value of c from different initializations, and evaluate the soft indices based on these partitions. The corresponding clustering parameters are specified in Table 6. The architecture for fuzzy clustering and evaluation is illustrated in Fig. 9.

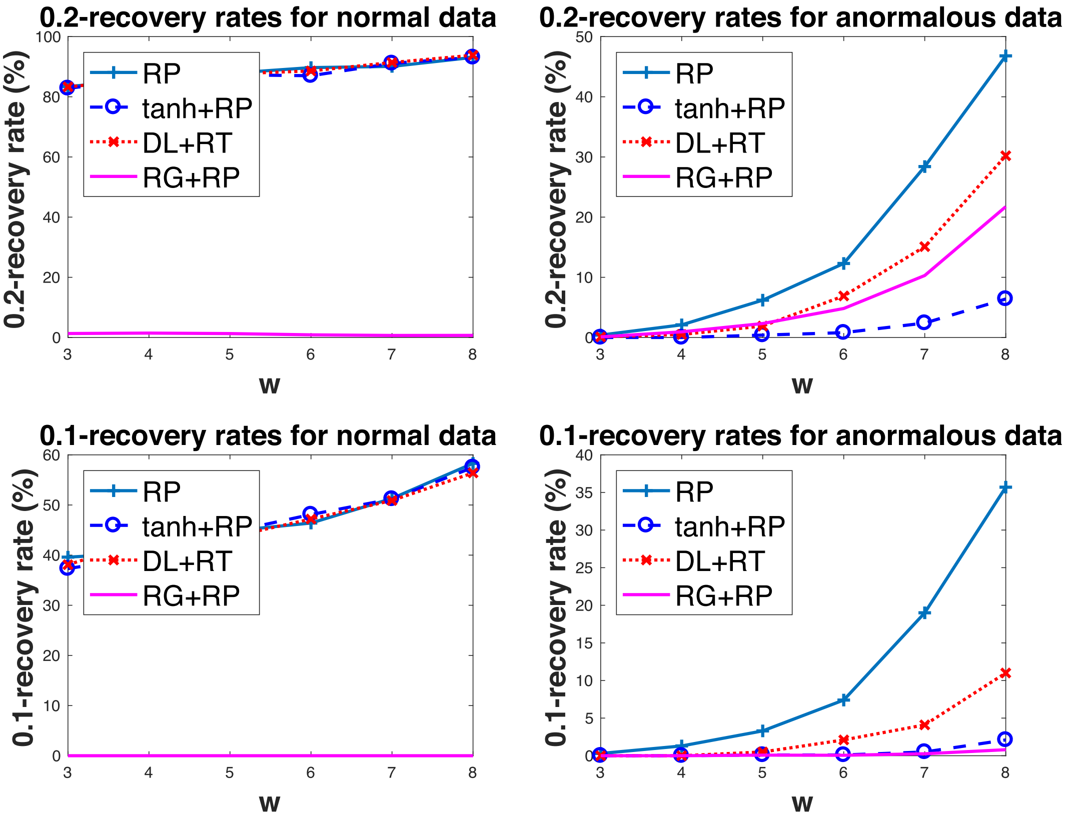


Fig. 6. Recovery rates of MAP estimation attacks against the evaluated schemes, on $w \times 1000$ data projected from 15×1000 normalised Laplace-distributed data (zero mean, unity scale).

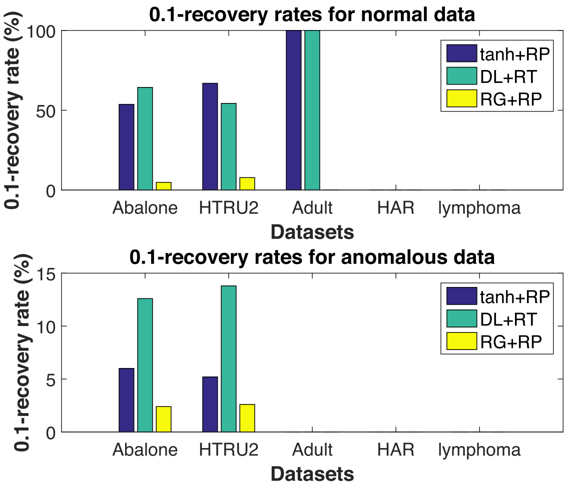


Fig. 7. 0.1-recovery rates of the MAP estimation attack against the evaluated schemes, on various datasets. The rank of the perturbation matrix, w, is set as $\lfloor (n+1)/2 \rfloor$, where n is the number of features. Note zero recovery rates in many cases.

6.2. *Evaluation of fuzzy clustering using cluster validity indices (CVIs)*

For evaluating clustering accuracy, Anderson et al. [43] proposed generalizations of hard *cluster validity indices* (CVIs) for use with soft partitions. Let $\mathbf{U}$ and $\mathbf{V}$ represent two partitions: $\mathbf{U}$ is any c-partition of N objects, and $\mathbf{V}$ is any r-partition of the N objects. Note that in general, $c \neq r$. The *contingency table* is the product $\mathbf{N} = \mathbf{U}\mathbf{V}^{\top}$ as shown in Table 7, where $n_{i\boldsymbol{\cdot}}$ is the row sum of row i of $\mathbf{N}$. In the soft clustering setting, $n_{i\boldsymbol{\cdot}}$ can be regarded as the probability that the n points belong to the ith cluster, n_{ij} represents the joint probability that a point belongs to clusters U_i and V_j. The numbers of pairs of shared objects between $\mathbf{U}$ and $\mathbf{V}$ are divided into four

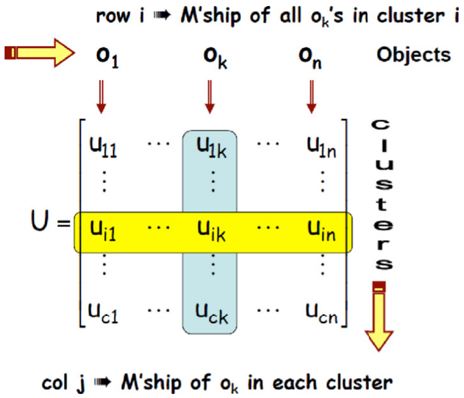


Fig. 8. A fuzzy partition matrix, where rows represent clusters, and columns represent objects.

Table 6
FCM parameters.

Parameters	Value
Exponent for the partition matrix **U**	2
Maximum number of iterations	100
Minimum amount of improvement	1e-5
Number of clusters	$[c_{true} - \frac{ctrue}{2}, c_{true} + \frac{ctrue}{2}]$
Number of initializations	100

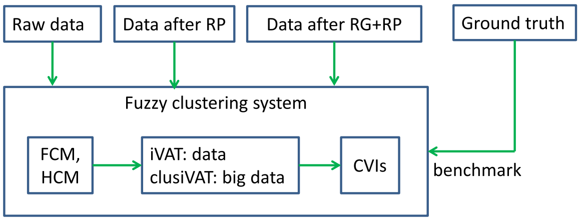


Fig. 9. Our fuzzy clustering and evaluation architecture.

Table 7
Contingency table for two partitions **U** and **V**.

		Partition **V**	
		V_j = row j of **V**	
	Class	$V_1\ V_2 \cdots V_r$	Sums
Partition **U** U_i = row i of **U**	U_1 U_2 U_3 $\vdots$ U_c	$\mathbf{N} = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1r} \\ n_{21} & n_{22} & \cdots & n_{2r} \\ n_{31} & n_{32} & \cdots & n_{3r} \\ \cdots & \cdots & \ddots & \cdots \\ n_{c1} & n_{c2} & \cdots & n_{cr} \end{bmatrix} = \mathbf{UV}^T$	$n_{1\cdot}$ $n_{2\cdot}$ $n_{3\cdot}$ $\vdots n_{c\cdot}$
	Sums	$n_{\cdot1}n_{\cdot2} \ldots n_{\cdot r}$	$n_{\cdot\cdot} = N$

groups:

- k_{11}, the number of pairs that are in the same clusters in both **U** and **V**;
- k_{00}, the number of pairs that are in different clusters in both **U** and **V**;
- k_{10}, the number of pairs that are in the same clusters in **U** but in different clusters in **V**; and
- k_{01}, the number of pairs that are in different clusters in **U** but in the same clusters in **V**.

The sum $k_{11} + k_{00}$ is interpreted as the total number of agreements between **U** and **V**, and the sum $k_{10} + k_{01}$ is the total number of disagreements. Based on the contingency table, information entropy, joint entropy and mutual information for soft clusterings are redefined [44] as:

Table 8
Cluster validity indices.

Indices	Expression	Range	Objective
RI	$\frac{k_{11}+k_{00}}{k_{11}+k_{10}+k_{01}+k_{00}}$	[0, 1]	max
ARI	$\frac{k_{11}-\frac{(k_{11}+k_{10})(k_{11}+k_{01})}{k_{11}+k_{10}+k_{01}+k_{00}}}{\frac{(k_{11}+k_{10})+(k_{11}+k_{01})}{2}-\frac{(k_{11}+k_{10})(k_{11}+k_{01})}{k_{11}+k_{10}+k_{01}+k_{00}}}$	[−1, 1]	max
MI	$I(\mathbf{U}, \mathbf{V})$	[0, min(H(**U**), H(**V**))]	max
NMI$_{\{sqrt\}}$	$\frac{I(\mathbf{U},\mathbf{V})}{\sqrt{H(\mathbf{U})H(\mathbf{V})}}$	[0, 1]	max
VI	$H(\mathbf{U}) + H(\mathbf{V}) - 2MI$	[0, log n]	min
NVI	$\frac{VI}{H(\mathbf{U},\mathbf{V})}$	[0, 1]	min
JVI	$1-\frac{MI}{\max(H(\mathbf{U}),H(\mathbf{V}))}$	[0, 1]	min

$$H(\mathbf{U}) = -\sum_{i=1}^{r} \frac{n_{i\bullet}}{N} \log \frac{n_{i\bullet}}{N}, \tag{10}$$

$$H(\mathbf{U}, \mathbf{V}) = -\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}}{N} \log \frac{n_{ij}}{N}, \tag{11}$$

$$I(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{n_{i\bullet} n_{\bullet j}/N^2}. \tag{12}$$

Here, we use two well-known non-information-theoretic indices and five information-theoretic indices as accuracy indices for comparing fuzzy clustering results between the raw data and perturbed data, with the hope that the results of clustering based on the perturbed data are close to the corresponding results using the original data. We also take the ground truth as a basis for evaluating CVIs. The non-information-theoretic indices are:

- *Rand index* (RI) proposed by W.M. Rand in 1971 [45]: An RI takes a value in [0, 1], which is 1 when the partitions are identical, but 0 when there is total disagreement between matching pairs.
- *Adjusted Rand Index* (ARI) proposed by Hubert and Arabie in 1985 [46]: An ARI peaks at 1, but its minimum can be less than 0.

The indices above assess cluster quality in terms of pair-based matches between two partitions. The remaining five information-theoretic indices are:

- *Mutual information* (MI)
- *Square root normalised mutual information* (NMI$_{sqrt}$)
- *Variation of Information* (VI)
- *Normalised VI* (NVI)
- *VI based on normalised Jaccard distance* (JVI).

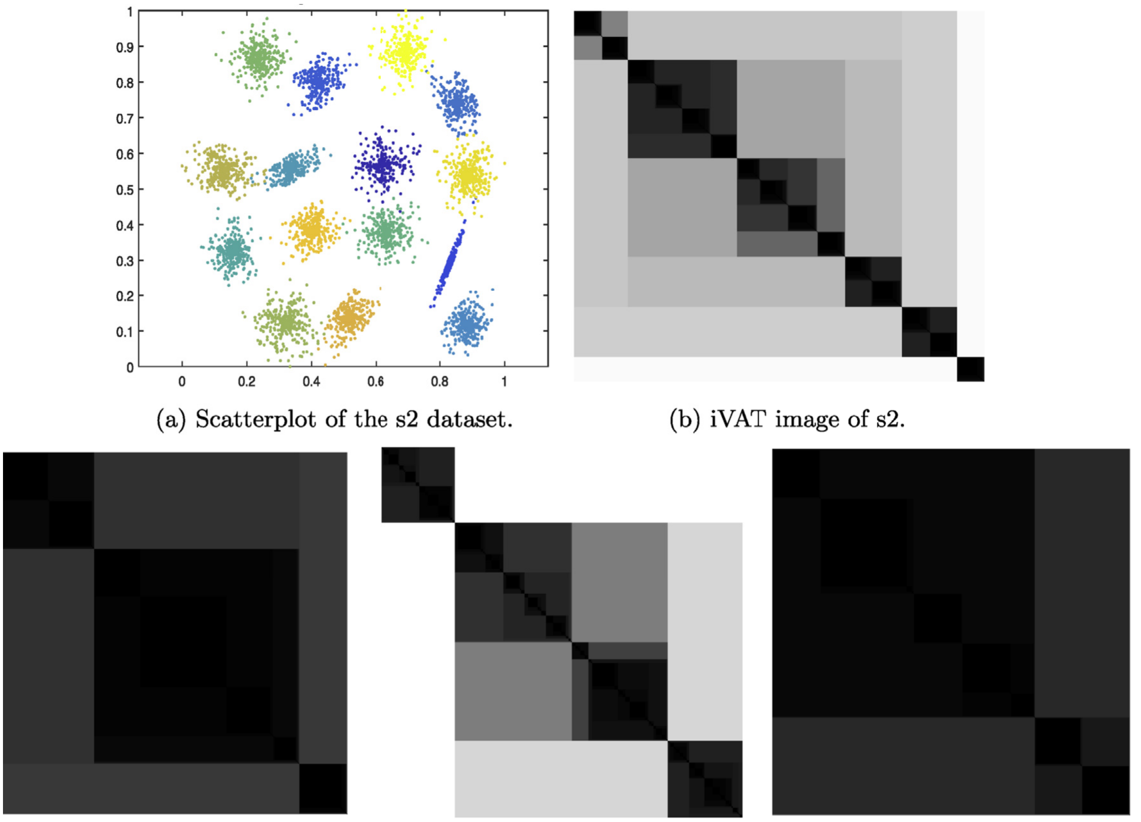
All seven indices shown in Table 8 are called external CVIs because they use a crisp ground truth partition of the data as a reference matrix.

6.3. Cluster accuracy evaluation

We evaluate the proposed scheme using different data sizes and dimensions. Experiments are conducted on three low-dimensional synthetic datasets ranging in size from quite small to medium, and five real datasets from the UCI Machine Learning Repository. For simulations, we randomly generate synthetic datasets s1, s2 and s3 with 1000, 5000, 100000 records respectively. For example, s2 is

Table 9
Experimental datasets.

Datasets	#records(m)	#classes(c)	Upspace dimension(n)	Downspace dimension(w)
s1	1000	4	2	2
s2	5000	15	2	2
s3	100000	4	2	2
Iris	150	3	4	2
HTRU2	17898	2	8	4
colonTumor	62	2	2000	1000
lymphoma	45	2	4026	2013
DrivFace	606	3	6400	3200



(a) Scatterplot of the s2 dataset.

(b) iVAT image of s2.

(c) iVAT image of s2 after RG.

(d) iVAT image of s2 after RP.

(e) iVAT image of s2 after RG+RP.

Fig. 10. Scatterplot and iVAT images for the s2 dataset.

a 2D dataset with $m = 5000$ samples, as shown in Fig. 11(a), it forms $c = 15$ synthetic Gaussian clusters [47]. The evaluation datasets are listed in Table 9. The effect of applying RP and RG + RP can be observed from the various visualisation and numerical evaluations.

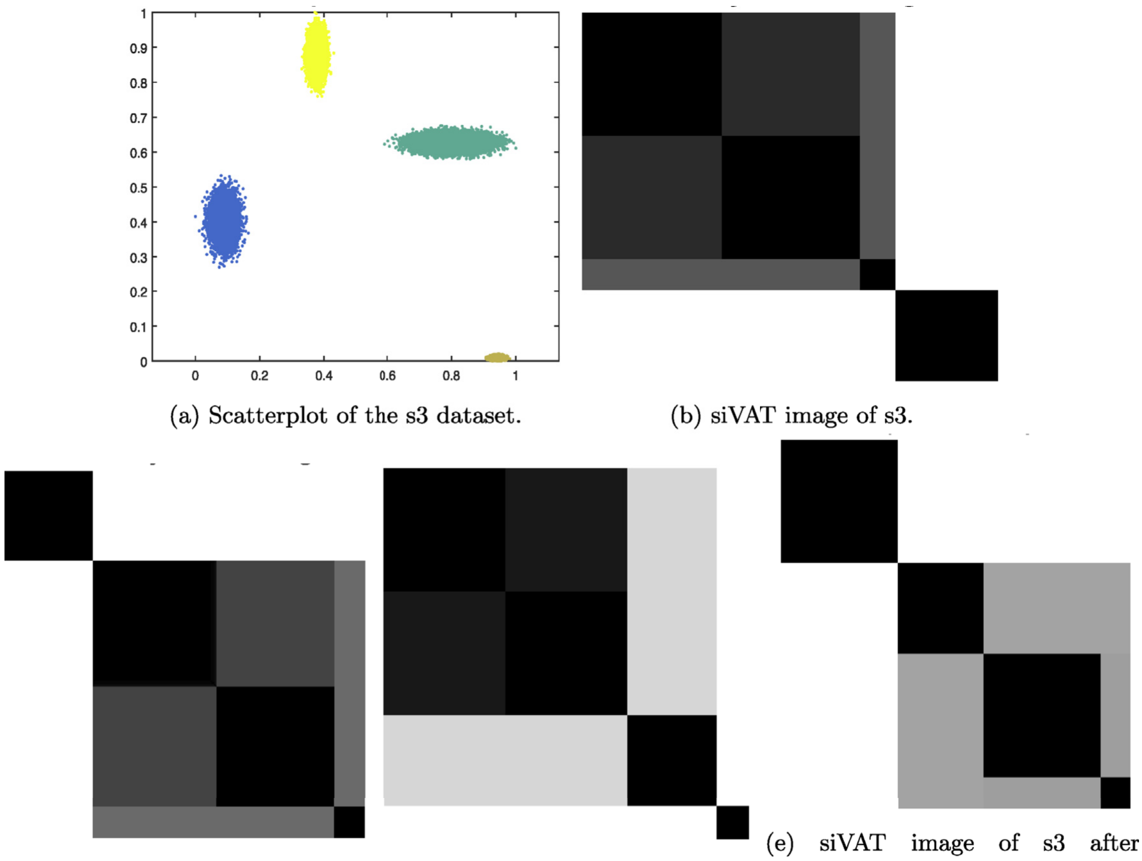
6.3.1. Visualisation results

In order to visually illustrate how the two-stage perturbation affects the data distribution, we use visual assessment tools to inspect the data at each stage of the process. The *Visual Assessment of Tendency* (VAT) algorithm [48] addresses the question of clustering tendency by reordering the dissimilarity matrix **D** to obtain $\mathbf{D}^*$ so that different clusters are displayed as dark blocks along the diagonal of the image of $\mathbf{D}^*$. While VAT provides a useful estimate of the number of clusters in a dataset, a much sharper reordered diagonal matrix image can be obtained using *improved VAT* (iVAT) [49,50]. The iVAT algorithm uses a graph-theoretic distance transform, as subsequently defined, to improve the effectiveness of VAT for cases where VAT fails to accurately show the cluster tendency. Given the dissimilarity matrix **D**, iVAT

- transforms **D** into $\mathbf{D}^{'}$, where the (i, j)-th entry of $\mathbf{D}^{'}$ is given by the minimum of the maximum path costs from object i to object j, and the cost of a path is measured by the maximum edge weight along the path;
- and applies iVAT to the transformed dissimilarity matrix $\mathbf{D}'$.

However, while VAT and iVAT work well for small datasets, they suffer from resolution and memory constraints. To overcome this limitation, Hathaway et al. [51] introduced *scalable VAT* (sVAT), which works by sampling a big dataset and then constructing a VAT image of the sample. Given a dataset, $\mathbf{X} = [\vec{x}_1\ \vec{x}_2\ \cdots\ \vec{x}_N]$, sVAT finds a small distance matrix of **X**, say $\mathbf{D}_{M\times M}$, where M is a “VAT-sized” fraction of N. *Scalable iVAT* (siVAT) is similar to sVAT, except it applies iVAT after the sampling step. clusiVAT is an extension of siVAT for big data clustering [52].

For experimentation, we apply iVAT to small datasets ($m \le 5000$) and clusiVAT to big datasets ($m > 5000$). To get a better idea of the overall process, Fig. 10 shows iVAT images of the s2 dataset at each stage of the process. The structure of the original s2 data is seen by iVAT in Fig. 10(b). Fig. 10(c) shows that the cluster structure in the data has been altered by the RG function (first-stage nonlinear transformation). Fig. 10(d) shows that after only RP (second-stage linear projection), the cluster structure in the data stays stable, although not exactly the same, which indicates that the second-stage RP of our two-stage transformation method poses minor accuracy loss, as the probabilistic nature of RP results in a small number of noisy or outlying data points. In addition, three clusters



(a) Scatterplot of the s3 dataset. (b) siVAT image of s3.

(c) siVAT image of s3 after RG. (d) siVAT image of s3 after RP. (e) siVAT image of s3 after RG+RP.

Fig. 11. Scatterplot and iVAT images for the s3 dataset.

can be seen in the iVAT image in Fig. 10(e), which is similar to Fig. 10(c), indicating they are mainly contributed by RG. A reasonable explanation for the performance degradation for the s2 dataset is that there exists 15 small clusters, after RG and RG + RP, many small clusters are merged into one large cluster.

Fig. 11 shows siVAT images of the big dataset s3 at each stage of the process. The structure of the original s3 dataset is seen by siVAT in Fig. 11(b). Fig. 11(c) and (d) show that both the first-stage RG and the second-stage RP do not cause any serious clustering accuracy loss compared with the results on the original data. Consistently, cluster structure in the data has not been altered by the two-stage transformation, as reflected in the improved clarity and contrast in the block structure in Fig. 11(e). The siVAT images of s3 indicate that RG and RG + RP do not pose serious accuracy loss, likely because of the clear pattern of s3 with four classes.

Figs. 10(d) and 11(d) suggest that the second-stage RP only pose limited accuracy loss, which verifies that random projection preserves distance-related statistics with high probability. On the other hand, it is not a great surprise that the loss of information in the overall process appears to be much greater in Fig. 10 than in Fig. 11. We conjecture that this is due to the proximity of clusters in Fig. 10(a); the clusters in Fig. 11(a) are much more compact and well separated.

6.3.2. *Numerical results*

For numerical evaluation of distance preservation, we use inner product and Euclidean distance to measure the similarities and distances between data vectors respectively, and evaluate the amount of distortion/difference by the Root Mean Squared Error (RMSE). For two-dimensional synthetic datasets, random projection (from $\mathbb{R}^2$ to $\mathbb{R}^2$) does not reduce the dimension. For high-dimensional datasets, we reduce the original dimension to half of their original dimensions by using a ($\lfloor (n + 1)/2 \rfloor \times n$) RP matrix to resist ICA attacks. We normalize all the data vectors and evaluate the amount of distortion caused by different stages by comparing:

- the column-wise inner product of these two data sets before and after row-wise random projection;
- the column-wise inner product of these two data sets before and after two-stage perturbation;
- the Euclidean distance between pairs of data points before and after row-wise random projection;
- the Euclidean distance between pairs of data points before and after two-stage perturbation.

Furthermore, we evaluate our scheme in terms of the preservation of crisp clusters by repeating the tests above with the *hard c-*

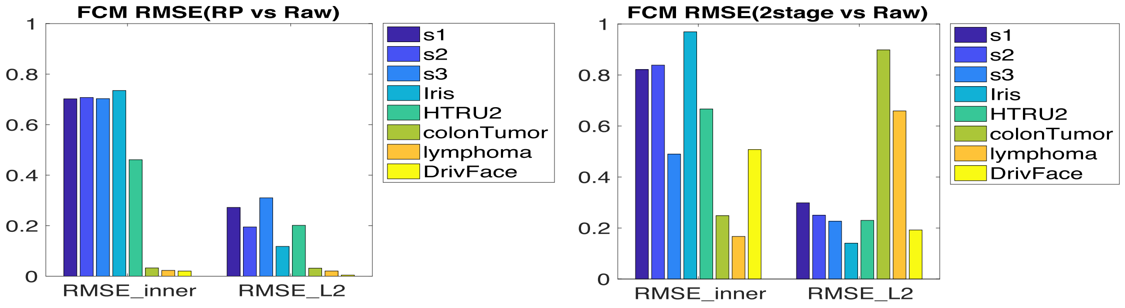


Fig. 12. Results for FCM: RMSE of inner product and Euclidean distance between RP-perturbed data and raw data, and between two-stage perturbed data and raw data.

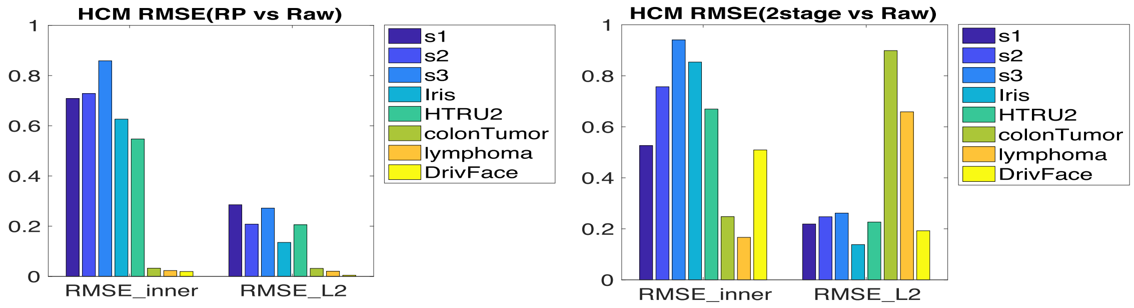


Fig. 13. Results for HCM: RMSE of inner product and Euclidean distance between RP-perturbed data and raw data, and between two-stage perturbed data and raw data.

means (HCM) clustering algorithm, which is the limiting case of FCM when the fuzzy partition matrix exponent is 1.

The RMSE results for inner product and Euclidean distance between RP-perturbed data and raw data, and between two-stage-perturbed data and raw data are shown in Figs. 12 and 13, for FCM and HCM respectively. All RMSE results are average values taken over 100 randomised simulation runs.

Similarly, for numerical evaluation of clustering accuracy, we compare the accuracy results based on raw data and perturbed data by defining accuracy ratio as shown in Eq. (13). Accuracy (*accuracy_perturbed* and *accuracy_raw*) can be computed by comparing the predicted cluster assignments and true labels (ground truth labels for the synthetic datasets and class labels for the real-world datasets). However, accuracy ratio directly indicates how perturbation on the raw data affects the algorithm's ability to identify classes comparing with the corresponding results based on raw data. In particular, the closer the accuracy ratio approaches to 1, the more similar the results of clustering based on the perturbed data to the corresponding results using the original data. When accuracy ratio equals 1, it means the perturbed data can produce the same patterns that can be extracted from the raw data.

$$accuracy\_ratio = \frac{accuracy\_perturbed}{accuracy\_raw} \tag{13}$$

The accuracy ratio results between RP-perturbed and raw data, and between two-stage perturbed data and raw data are shown in

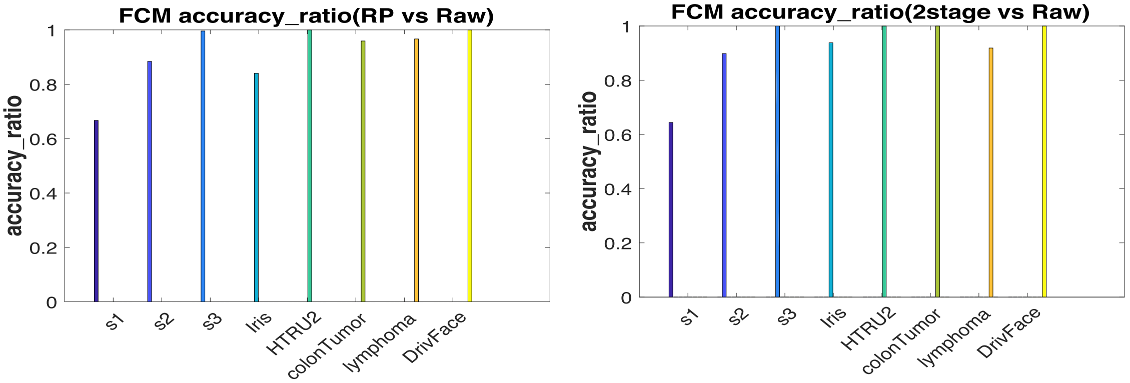


Fig. 14. Results for FCM: Clustering accuracy ratio between RP-perturbed data and raw data, and between two-stage perturbed data and raw data.

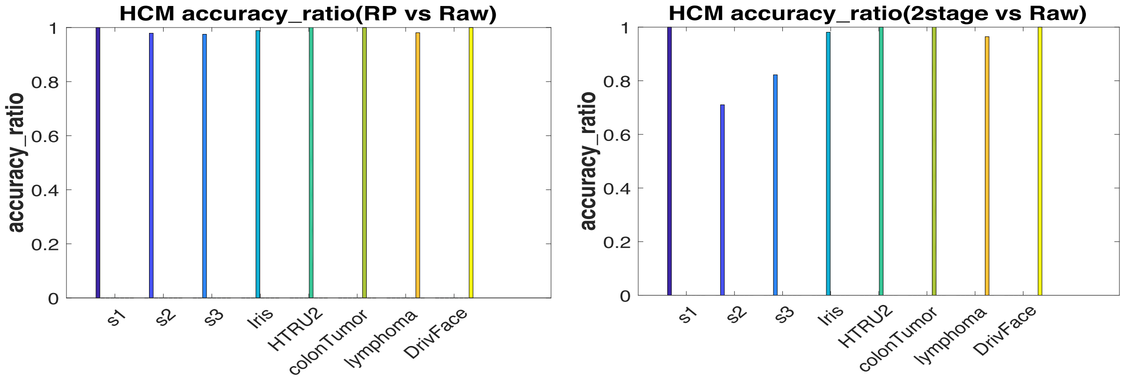


Fig. 15. Results for HCM: Clustering accuracy ratio between RP-perturbed data and raw data, and between two-stage perturbed data and raw data.

Figs. 14 and 15, for FCM and HCM respectively. All the results are average values taken over 100 randomised simulation runs.

The RMSE results indicate that similarities and Euclidean distances are well preserved even when the data dimensionality is reduced by half through either RP or RG + RP. Furthermore, the Euclidean distances between data vectors would probably have been preserved better. Consistent with the RMSE results, the clustering accuracy ratio approaches to 1 in most cases, i.e., results of clustering based on the perturbed data are close to the corresponding results using the raw data, demonstrating that clustering results based on RP-perturbed or two-stage-perturbed data are comparable to clustering results based on raw data — this confirms the utility of our privacy-preserving scheme when used with either FCM or HCM. By comparing Fig. 14(a) with Figs. 14(b), and Fig. 15(a) with Fig. 15(b), RG + RP is found to perform similarly to RP alone, implying that RG does not reduce the clustering accuracy significantly when using FCM or HCM as the basic clustering algorithm. A possible explanation for this is that RG is designed to be nearly consistent in trend with the original data. Due to Proposition 1, comparable clustering accuracy is expected when any other distanced-based clustering algorithm is used.

To provide a better insight into how the CVIs in Table 8 perform, the *CVI_ratio* for an index is computed as the CVI value of the perturbed data divided by the CVI value of the raw data, both using the ground truth labels for the synthetic datasets and true class labels for the real-world datasets as the benchmark partition. The expression for *CVI_ratio* is shown in Eq. (14). The *CVI_ratio* results between RP-perturbed data and raw data, and between two-stage perturbed data and raw data are shown in Figs. 16 and 17, for FCM and HCM respectively. All the results are average values taken over 100 randomised simulation runs.

$$CVI\_ratio = \frac{CVI\_perturbed}{CVI\_raw} \tag{14}$$

When the CVI generated from the perturbed data equals to the CVI generated from the raw data, *CVI_ratio* achieves the best result which corresponds to 1, as highlighted in red line in Figs. 16 and 17. The graph shows that RI, VI, NVI and JVI perform similarly well, and can achieve a *CVI_ratio* of 1 in the best case, except for RP perturbed s3 dataset in HCM. In addition, in most cases, the information-theoretic soft indices (VI, NVI and JVI) behave slightly better than the non-information-theoretic indices (ARI, RI), except MI and NMIsqrt. We hypothesize that this is because the information-theoretic measures are good at distinguishing clusters with non-linear relationships [53]. In contrast, MI and NMIsqrt perform not as well as the other indices in most cases. We hypothesize the reason for this is that MI monotonically increases with the number of clusters c [53]. Hence, MI tends to favour clusterings with more clusters.

In addition, we use the *consensus index* (CI) [54] to evaluate different CVIs. CI is based on the idea of consensus clustering that aims to produce a robust and high-quality representative clustering by considering a set of partitions generated from the same dataset. The definition of CI follows: suppose a set of L clustering solutions (crisp or soft), denoted $S_c = \{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_L\}$, have been

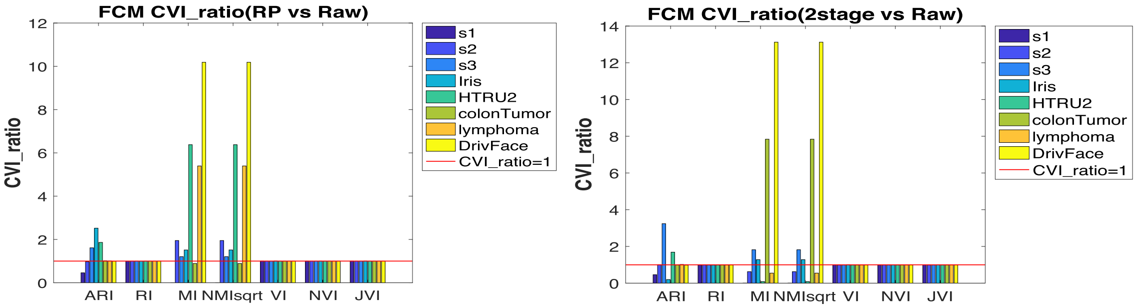


Fig. 16. Results for FCM: CVI ratio between RP-perturbed data and raw data, and between two-stage perturbed data and raw data.

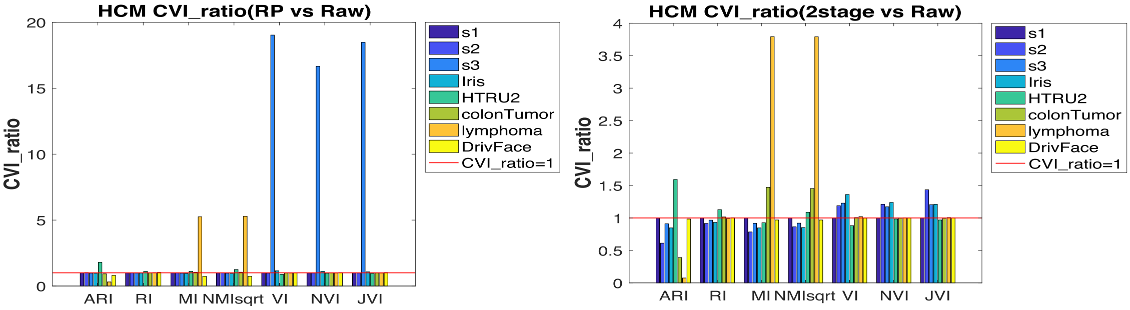


Fig. 17. Results for HCM: CVI ratio between RP-perturbed data and raw data, and between two-stage perturbed data and raw data.

Table 10
True classes vs clusters corresponding to the best CI value for FCM/HCM (F/H).

Datasets	class	ARI (F/H)	RI (F/H)	MI (F/H)	NMI$_{sqrt}$ (F/H)	VI (F/H)	NVI (F/H)	JVI (F/H)
s1	4	4/4	6/4	4/6	4/4	2/4	4/4	4/4
s2	15	2/16	16/16	4/16	4/5	2/2	4/16	4/16
s3	4	3/4	6/4	4/6	4/4	2/2	4/4	3/4
Iris	3	3/3	4/3	4/3	4/3	2/2	3/3	3/3
HTRU2	2	2/2	2/2	3/3	2/3	2/2	2/3	2/2
colonTumor	2	2/3	2/3	2/3	2/3	2/2	2/3	2/3
lymphoma	2	2/2	2/2	3/3	2/2	2/2	2/2	2/2
DrivFace	3	2/2	2/2	4/3	4/3	2/2	4/2	2/2

generated, each with c clusters. The CI of S_c is defined as

$$\mathrm{CI}(S_c) = \frac{\sum_{i<j} \mathrm{AM}(\mathbf{U}_i, \mathbf{U}_j)}{L(L-1)/2}, \tag{15}$$

where AM is an *agreement measure*, i.e., a suitable clustering similarity index. For the agreement measure, we use the CVIs in Table 8.

For implementation, 100 partitions are generated by the FCM algorithm for each value of c. Seven CVIs are computed on the candidate partitions for each dataset. The clusters associated with the best CI value in FCM/HCM for different CVIs are listed in Table 10. The table shows that for both FCM and HCM, all the CVIs tend to perform best when the formed clusters fall around the true classes, implying the capability of these CVIs to find the true patterns, especially for datasets with clear patterns or less classes, such as HTRU2, colonTumor, lymphoma and DrivFace. However, it is obvious that the number of true classes, may not correspond to the best CI value (number of apparently best clusters), found by a computational clustering algorithm. A possible reason is that the clustering algorithm fails to detect the underlying substructure of the data, rather than the CVI fails [44]. Furthermore, it is entirely possible that the clustering algorithm fails to detect the underlying complex substructure of the data. For example, an algorithm that is designed to detect spherical clusters, cannot reliably detect elongated clusters [44].

Moreover, we also observe that for the second-stage perturbation, the RMSE and accuracy ratio results associated with RP are better than the results associated with uniform random transformation (by around 5%–10%, but no plots provided here). This is consistent with Proposition 1.

7. Conclusion and future work

We presented a two-stage privacy-preserving collaborative fuzzy clustering scheme. The randomisation-based scheme perturbs data in two stages: the first, nonlinear stage thwarts Bayesian estimation attack and mitigates collusion attack, whereas the second, linear stage resists independent component analysis attack.

The nonlinear function is designed to condition the pdf of the perturbed data to protect both anomalous and normal data records. Our analysis on several benchmark datasets reveals that the two-stage transformation, RG + RP, preserves privacy of both normal and anomalous data, and delivers the lowest recovery rates for all the benchmark datasets, outperforming techniques based on other non-linear schemes, such as tanh + RP and DL + RT.

To evaluate distortion before and after perturbation, inner product and Euclidean distance between data vectors are numerically measured by RMSE. Similarly, the accuracy ratio results by comparing predicted cluster assignments generated from FCM and HCM on raw data and data perturbed by different stages are provided. The low RMSE and high accuracy ratio results manifest that our proposed privacy-preserving clustering method achieves comparable results with the non-private clustering method based on raw data, and it can be applied to both FCM and HCM. Furthermore, CVI ratio is calculated and CI is used to evaluate the best candidates generated by FCM and HCM based on seven different CVIs. It is observed that with the exception of MI and NMIsqrt, all the indices perform reasonably well for evaluation, and all the CVIs tend to perform best when the formed clusters fall around the true classes, especially for dataset with clear patterns. Consequently, RG + RP appears to be a good trade-off between accuracy and privacy.

The next step from this work is to consider other kinds of attacks and different transformation matrices for different clustering mechanisms.

Acknowledgments

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

References

[1] D. Estrin, K.M. Chandy, R.M. Young, L. Smarr, A. Odlyzko, D. Clark, V. Reding, T. Ishida, S. Sharma, V.G. Cerf, et al., Participatory sensing: applications and architecture [internet predictions], IEEE Internet Comput. 14 (1) (2010) 12–42.
[2] B. Liu, Y. Jiang, F. Sha, R. Govindan, Cloud-enabled privacy-preserving collaborative learning for mobile sensing, Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems, ACM, 2012, pp. 57–70.
[3] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Publishers, 1981.
[4] R. Cramer, I.B. Damgøard, J.B. Nielsen, Secure Multiparty Computation and Secret Sharing, Cambridge University Press, 2015 Cambridge Books Online https://doi.org/10.1017/CBO9781107337756.
[5] Z. Huang, W. Du, B. Chen, Deriving private information from randomized data, Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, ACM, 2005, pp. 37–48.
[6] K. Chen, G. Sun, L. Liu, Towards attack-resilient geometric data perturbation, Proceedings of the 2007 SIAM International Conference on Data Mining, SIAM, 2007, pp. 78–89.
[7] Y. Sang, H. Shen, H. Tian, Effective reconstruction of data perturbed by random projections, IEEE Trans. Comput. 61 (1) (2012) 101–117.
[8] C.R. Giannella, K. Liu, H. Kargupta, Breaching euclidean distance-preserving data perturbation using few known inputs, Data Knowl. Eng. 83 (2013) 93–110.
[9] L. Lyu, Y.W. Law, S.M. Erfani, C. Leckie, M. Palaniswami, An improved scheme for privacy-preserving collaborative anomaly detection, 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), IEEE, 2016, pp. 1–6.
[10] K. Liu, H. Kargupta, J. Ryan, Random projection-based multiplicative data perturbation for privacy preserving distributed data mining, Knowl. Data Eng. IEEE Trans. 18 (1) (2006) 92–106.
[11] Y. Lei, J.C. Bezdek, J. Chan, N.X. Vinh, S. Romano, J. Bailey, Extending information-theoretic validity indices for fuzzy clustering, IEEE Trans. Fuzzy Syst. 25 (4) (2017) 1013–1018.
[12] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, Theory of Cryptography Conference, Springer, 2006, pp. 265–284.
[13] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, M. Naor, Our data, ourselves: privacy via distributed noise generation, Annual International Conference on the Theory and Applications of Cryptographic Techniques, Springer, 2006, pp. 486–503.
[14] B. Chor, S. Goldwasser, S. Micali, B. Awerbuch, Verifiable secret sharing and achieving simultaneity in the presence of faults, Foundations of Computer Science, 1985., 26th Annual Symposium on, IEEE, 1985, pp. 383–395.
[15] E. Shi, H. Chan, E. Rieffel, R. Chow, D. Song, Privacy-preserving aggregation of time-series data, Annual Network & Distributed System Security Symposium (NDSS), Internet Society, 2011.
[16] G. Ács, C. Castelluccia, I have a dream!(differentially private smart metering), International Workshop on Information Hiding, Springer, 2011, pp. 118–132.
[17] R. Agrawal, R. Srikant, Privacy-preserving data mining, Proceedings of the ACM SIGMOD Conference on Management of Data, ACM, 2000, pp. 439–450.
[18] R.K. Ganti, N. Pham, Y.-E. Tsai, T.F. Abdelzaher, Poolview: stream privacy for grassroots participatory sensing, Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems, ACM, 2008, pp. 281–294.
[19] F. Zhang, L. He, W. He, X. Liu, Data perturbation with state-dependent noise for participatory sensing, INFOCOM, 2012 Proceedings IEEE, IEEE, 2012, pp. 2246–2254.
[20] K. Chen, L. Liu, Privacy preserving data classification with rotation perturbation, Data Mining, Fifth IEEE International Conference on, IEEE, 2005, p. 4.
[21] W.B. Johnson, J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, Contemp. Math. 26 (1984) 189–206.
[22] S. Kaski, Dimensionality reduction by random mapping: fast similarity computation for clustering, Neural Networks Proceedings, 1998. Ieee World Congress on Computational Intelligence. The 1998 Ieee International Joint Conference on, vol. 1, IEEE, 1998, pp. 413–418.
[23] P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, ACM, 1998, pp. 604–613.
[24] E. Bingham, H. Mannila, Random projection in dimensionality reduction: applications to image and text data, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2001, pp. 245–250.
[25] D. Achlioptas, Database-friendly random projections: johnson-lindenstrauss with binary coins, J. Comput. Syst. Sci. 66 (4) (2003) 671–687.
[26] A. Dasgupta, R. Kumar, T. Sarlós, A sparse johnson: lindenstrauss transform, Proceedings of the Forty-second ACM Symposium on Theory of Computing, ACM, 2010, pp. 341–350.
[27] D.M. Kane, J. Nelson, Sparser johnson-lindenstrauss transforms, J. ACM 61 (1) (2014) 4.
[28] K. Liu, C. Giannella, H. Kargupta, A survey of attack techniques on privacy-preserving data perturbation methods, Privacy-preserving Data Mining, Springer, 2008, pp. 359–381.
[29] S.M. Erfani, Y.W. Law, S. Karunasekera, C.A. Leckie, M. Palaniswami, Privacy-preserving collaborative anomaly detection for participatory sensing, Pacific-asia Conference on Knowledge Discovery and Data Mining, Springer, 2014, pp. 581–593.
[30] O.L. Mangasarian, E.W. Wild, G.M. Fung, Privacy-preserving classification of vertically partitioned data via random kernels, ACM Trans. Knowl. Discov. Data 2 (3) (2008) 12:1–12:16.
[31] K. Bhaduri, M.D. Stefanski, A.N. Srivastava, Privacy-preserving outlier detection through random nonlinear data distortion, Syst. Man, and Cybern. Part B: Cybern. IEEE Trans. 41 (1) (2011) 260–272.
[32] A. Cornuejols, C. Wemmert, P. Gançarski, Y. Bennani, Collaborative clustering: why, when, what and how, Inf. Fusion 39 (2018) 81–95.
[33] G. Forestier, P. Gancarski, C. Wemmert, Collaborative clustering with background knowledge, Data Knowl. Eng. 69 (2) (2010) 211–228.
[34] G. Cleuziou, M. Exbrayat, L. Martin, J.-H. Sublemontier, Cofkm: a centralized method for multiple-view clustering, Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on, IEEE, 2009, pp. 752–757.
[35] S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estrin, M. Hansen, Image browsing, processing, and clustering for participatory sensing: lessons from a dietsense prototype, Proceedings of the 4th Workshop on Embedded Networked Sensors, ACM, 2007, pp. 13–17.
[36] S. Merugu, J. Ghosh, Privacy-preserving distributed clustering using generative models, Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, IEEE, 2003, pp. 211–218.
[37] W. Pedrycz, Collaborative fuzzy clustering, Pattern Recogn. Lett. 23 (14) (2002) 1675–1686.
[38] L.F. Coletta, L. Vendramin, E.R. Hruschka, R.J. Campello, W. Pedrycz, Collaborative fuzzy clustering algorithms: some refinements and design guidelines, IEEE Trans. Fuzzy Syst. 20 (3) (2012) 444–462.
[39] J. Zhou, C.P. Chen, L. Chen, H.-X. Li, A collaborative fuzzy clustering algorithm in distributed network environments, IEEE Trans. Fuzzy Syst. 22 (6) (2014) 1443–1456.

[40] S. Rahimi, M. Zargham, A. Thakre, D. Chhillar, A parallel fuzzy c-mean algorithm for image segmentation, Fuzzy Information, 2004. Processing NAFIPS'04. IEEE Annual Meeting of the, vol. 1, IEEE, 2004, pp. 234–237.

[41] K. Liu, C. Giannella, H. Kargupta, An attacker's view of distance preserving maps for privacy preserving data mining, Knowledge Discovery in Databases: PKDD 2006, Springer, 2006, pp. 297–308.

[42] K. Liu, Multiplicative Data Perturbation for Privacy Preserving Data Mining, Ph.D. thesis University of Maryland, Baltimore County, 2007.

[43] D.T. Anderson, J.C. Bezdek, M. Popescu, J.M. Keller, Comparing fuzzy, probabilistic, and possibilistic partitions, IEEE Trans. Fuzzy Syst. 18 (5) (2010) 906–918.

[44] Y. Lei, J.C. Bezdek, J. Chan, N.X. Vinh, S. Romano, J. Bailey, Generalized information theoretic cluster validity indices for soft clusterings, Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on, IEEE, 2014, pp. 24–31.

[45] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (336) (1971) 846–850.

[46] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1) (1985) 193–218.

[47] P. Fränti, O. Virmajoki, Iterative shrinking method for clustering problems, Pattern Recogn. 39 (5) (2006) 761–775.

[48] J.C. Bezdek, R.J. Hathaway, Vat: a tool for visual assessment of (cluster) tendency, Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on, vol. 3, IEEE, 2002, pp. 2225–2230.

[49] L. Wang, X. Geng, J. Bezdek, C. Leckie, R. Kotagiri, Enhanced visual analysis for cluster tendency assessment and data partitioning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1401–1414.

[50] T.C. Havens, J.C. Bezdek, An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm, IEEE Trans. Knowl. Data Eng. 24 (5) (2012) 813–822.

[51] R.J. Hathaway, J.C. Bezdek, J.M. Huband, Scalable visual assessment of cluster tendency for large data sets, Pattern Recogn. 39 (7) (2006) 1315–1324.

[52] D. Kumar, J.C. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie, T.C. Havens, A hybrid approach to clustering in big data, IEEE Trans. Cybern. 46 (10) (2016) 2372–2385.

[53] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: is a correction for chance necessary? Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 1073–1080.

[54] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance, J. Mach. Learn. Res. 11 (Oct) (2010) 2837–2854.