Artif Intell Rev (2015) 44:547–569
DOI 10.1007/s10462-015-9439-5

A survey: deriving private information from perturbed data

Burcu D. Okkalioglu[1] · Murat Okkalioglu[1] · Mehmet Koc[2] · Huseyin Polat[3]

Published online: 14 September 2015
© Springer Science+Business Media Dordrecht 2015

Abstract Privacy-preserving data mining has attracted the attention of a large number of researchers. Many data perturbation methods have been proposed to ensure individual privacy. Such methods seem to be successful in providing privacy and accuracy. On one hand, different methods are utilized to preserve privacy. On the other hand, various data reconstruction approaches have been proposed to derive private information from perturbed data. Thus, many researchers have been conducting various studies about data reconstruction methods and the resilience of data perturbation schemes. In this survey, we focus on data reconstruction methods due to their importance in privacy-preserving data mining. We provide a detailed review of the data reconstruction methods and the data perturbation schemes attacked by different data reconstruction techniques. We merge our review with the evaluation metrics and the data sets used in current attack techniques. Finally, we pose some open questions to provide a better understanding of these approaches and to guide future study.

Keywords Data reconstruction · Data perturbation · Privacy · Attack resilience · Spectral filtering

1 Introduction

Privacy-preserving data mining (PPDM) is a major subject in the field of data mining. It has attracted many researchers due to increasing concerns for the privacy issues of individuals. Thus, extracting useful information and uncovering hidden patterns from huge amounts of data without jeopardizing confidentiality are subjects that have receive an increasing amount

✉ Huseyin Polat
polath@anadolu.edu.tr

1 Computer Engineering Department, Yalova University, 77100 Yalova, Turkey

2 Electrical and Electronics Engineering Department, Bilecik Seyh Edebali University, 11210 Bilecik, Turkey

3 Computer Engineering Department, Anadolu University, 26470 Eskisehir, Turkey

of attention. The number of online transactions has rapidly increased with the unprecedented growth of the Internet. Users are more concerned about privacy, and they are reluctant to give away any piece of personal information. The hesitation to share confidential information guarantees that data mining approaches guard individual privacy. At this point, data perturbation methods play an important and critical role in maintaining confidentiality. Data perturbation techniques help individual users maintain privacy in data mining applications while providing accurate estimations with perturbed data. As privacy level increases, accuracy decreases due to augmented randomness. Hence, privacy and accuracy are usually considered conflicting goals. Data perturbation schemes should provide a good balance of privacy and accuracy.

Individual users seek to protect their private data by utilizing different data perturbation methods before sending their data to different data mining engines. Although such data perturbation techniques support privacy, their resilience against possible attacks should be discussed. It is possible that malicious service providers may want to derive private information from perturbed data. Hence, the research community investigates the vulnerabilities of the proposed data perturbation techniques and presents possible attack techniques that can exploit these vulnerabilities. These studies generally define data perturbation methods in detail and outline the possible ways that such methods can be exploited.

Attacking systems holding perturbed data with various data reconstruction schemes is an approach that has received an increasing amount of attention. Although there are various studies focusing on deriving information from perturbed data, few studies offer brief reviews of such works. In this study, we primarily investigate the additive and the multiplicative perturbation techniques and available data reconstruction methods for the data perturbed by these methods. Although we primarily focus on the additive and the multiplicative perturbation methods, we may cover other perturbation methods if they are included in the domain of applicability of the reconstruction techniques. This work also reviews the statistical analysis and evaluation of studies in the field to help researchers decide on analytical techniques for future study.

This paper is organized as follows. In Sect. 2, we briefly cover the related survey papers and their scopes while stating how our paper differs from them. We set our focus on the related studies deriving private information from perturbed data and present studies about data reconstruction techniques in Sect. 3. In Sect. 4, we give an introduction on current data perturbation methods and briefly express the studies that target these data perturbation schemes. We discuss the statistical analysis, the evaluation metrics, and the data sets used by these studies in Sect. 5; in Sect. 6, we mention open questions in this field. We conclude our survey with final thoughts and possible future directions in the last section.

2 Related surveys

Researchers devote special attention to deriving private information from perturbed data. Although there are novel studies that focus on finding confidential data from perturbed data, there are only a couple of survey papers covering a limited number of related studies.

In the literature, works by Liu (2007) and Guo (2007) are the foremost studies on the vulnerabilities of data perturbation methods. The authors have put a great deal of effort into defining the vulnerabilities of different data perturbation methods; their dissertations include contributions to the field and a review of different attacks. Although their studies play an important role in the field, they cover a limited number of studies through the year 2007.

Aggarwal and Yu (2008) survey randomization methods for PPDM. The authors briefly review the data reconstruction methods for randomization. In another work, Chen and Liu (2008) present a survey study on multiplicative data perturbation methods. In their study, the authors discuss different aspects of multiplicative data perturbation schemes and review the data reconstruction attacks used to attack such data perturbation techniques. Although these two surveys cover the data reconstruction methods for privacy preserving approaches, they do not particularly focus on attack techniques in PPDM, and their scope with respect to the attack techniques is limited.

Contrary to Aggarwal and Yu (2008) and Chen and Liu (2008), whose main concern is not data reconstruction methods for confidential data, Liu et al. (2008a) primarily focus on attack techniques for PPDM. In their study, the authors pay detailed attention to the attacks that derive private information from perturbed data in different PPDM applications. They discuss those attack techniques targeting additive and multiplicative data perturbation schemes while briefly addressing k-anonymization. However, their study covers attack techniques through the year 2008; and data reconstruction methods for perturbed data have received a good deal of attention from researchers since then. Our work addresses the efforts of researchers from early studies such as Agrawal and Srikant (2000) to current evaluation metrics and data sets for a better grasp of the proposed methods. We also state some possible open questions in this field to note deficiencies and guide possible future studies in our survey.

3 Data perturbation techniques

Data perturbation means transforming data into a different form. Data subjected to data perturbation will hereafter be called perturbed data. Data owners prefer data perturbation because they do not want to divulge their privacy. In PPDM applications, although data perturbation provides privacy for data owners, it should not neglect accuracy. Data perturbation techniques should preserve a good balance between privacy and accuracy because they conflict with each other. To address this issue, perturbed data must preserve the properties of the original data that will yield the same or similar outcomes when evaluated in PPDM models. In this section, we review different data perturbation methods used in PPDM applications.

Random perturbation or *randomization* is a widely used data perturbation scheme. Agrawal and Srikant (2000) first introduced a data perturbation scheme to the PPDM community. In their randomization scheme, a random number is added to an original data value. Assuming x_i is an original data value and r is a random value drawn from some distribution (uniform or Gaussian), $x_i + r$ instead of x_i will appear in the data set. Various randomization methods are investigated (Agrawal and Aggarwal 2001; Evfimievski et al. 2002; Rizvi and Haritsa 2002; Du and Zhan 2003). While the distribution of the random noises is known, some researchers investigate how to reconstruct the distribution of the original data (Agrawal and Srikant 2000; Agrawal and Aggarwal 2001; Domingo-Ferrer et al. 2004). Kargupta et al. (2003a, b, 2005) note that randomization may not hide sensitive information. They claim that additive data distortion preserves very little data privacy when there is strong correlation among data attributes. According to the empirical results presented by Huang et al. (2005), the higher the correlations among attributes, the more the original data can be reconstructed accurately. Zhang et al. (2006) challenge the additive data perturbation scheme by developing two data reconstruction methods based on k-means clustering and SVD. Their experiments indicate that a large amount of original data can be derived based on their proposed methods.

Multiplicative data perturbation techniques are generally based on exactly preserving (Oliveira and Zaïane 2003b; Chen and Liu 2005), *rotation perturbation*, or approximately preserving (Liu et al. 2006a), *random projection*, Euclidean distances between original columns and their counterparts in perturbed data. Original data is perturbed with $Y = MX$ in rotation perturbation, where Y is perturbed data, X is original data, and M is transformation matrix. Note that if M is considered to be an orthogonal mixing matrix, Euclidean distance is exactly preserved (Oliveira and Zaïane 2003b; Chen and Liu 2005), and this enables data miners to apply different algorithms utilizing Euclidean distance. On the other hand, reconstructing the original data X from the perturbed data Y is studied in various reconstruction methods for *rotation perturbation*. Liu et al. (2006b) and Giannella et al. (2013) work on privacy breach probability. Independent component analysis (ICA) definition is well suited for rotation perturbation, although it has some drawbacks. Different scholars (Liu et al. 2006a; Chen et al. 2007) analyze rotation perturbation in terms of ICA. Distance inference attacks targeting rotation perturbation are discussed by Chen et al. (2007) and Turgay et al. (2008). *Random projection* is based on a rectangular mixing matrix instead of the orthogonal one proposed in rotation perturbation (Liu et al. 2006a; Liu 2007). *Random projection* is constructed upon Johnson and Lindenstrauss (1984). It does not fully preserve Euclidean distance, but it preserves Euclidean distance approximately. Zhao et al. (2014) propose the l_1*-recovery* attack for random projection when the mixing matrix is leaking. Their proposed attack works when the original data is sparse.

Similar to a data projection-based data perturbation scheme, Guo and Wu (2007) define *arbitrary data projection*, where the transformation matrix can be any matrix so that Euclidean distance is not guaranteed to be preserved. Guo and Wu (2006b) also define *general linear transformation* perturbation that combines rotation perturbation and randomization. In other words, data are perturbed as $Y = MX + E$, where E is a noise matrix. Liu et al. (2008b) study privacy vulnerabilities for general perturbation with a Gaussian noise matrix having zero mean and arbitrary variance.

Random geometric perturbation is proposed by Chen et al. (2007) to enhance the resilience of *random perturbation* against *naïve estimation, ICA-based, rotation center*, and *distance inference* attacks. The authors analyze these attacks and propose an iterative algorithm to identify a transformation matrix maximizing the privacy guarantee (Chen et al. 2007; Chen and Liu 2005). To address *rotation center* and *distance inference* attacks, Chen et al. (2007) add random translation and a Gaussian noise matrix to the perturbation, respectively.

Randomized response techniques (RRTs) were first proposed by Warner (1965). His naïve approach was to identify the percentage of population having a sensitive attribute by asking a sensitive question with probability θ and another question whose answer is the opposite of a sensitive question with probability $1 - \theta$. Polat and Du (2006) propose a central server based scheme to disguise user ratings. Du and Zhan (2003) develop a decision tree based on the RRTs technique, while Guo and Wu (2009) apply RRTs for categorical data. Huang and Du (2008) tries to find an optimal perturbing matrix for RRTs.

Data sanitization is a process that helps data sharing among organizations while protecting sensitive knowledge in the form of sensitive association rules. Data sanitization is proposed by Atallah et al. (1999). The basic idea behind data sanitization is to transform the original database into the released database, also called the sanitized database, and to offer protection against disclosure of the sensitive data while maximizing data utility. There are a number of algorithms to sanitize the database (Saygin et al. 2001; Oliveira and Zaïane 2002, 2003a; Amiri 2007). Sramka et al. (2009) and Sramka (2010, 2012) propose data mining and fusion techniques to remove noise from the published sanitized data. The results show that non-

expert adversaries can remove a reasonable amount of added noise from the perturbed data without having any background information.

BLoom-then-flIP (BLIP) is a differentially private, which basically means that adding or removing a single row from a data set should not significantly alter the output, a structure that tries to ensure the privacy of user profiles (Alaggan et al. 2012). BLIP first calculates the Bloom filter of a profile and then flips Bloom filter output with a probability of ρ to create perturbed data. Kenthapadi et al. (2013) propose *privacy via Jonhson–Lindenstrauss Transform* (privacy-JLT). Privacy-JLT projects high dimensional vector representation into a lower dimensional space by applying the Jonhson and Lindenstrauss transform and then adding Gaussian noise to each bit in the lower dimension. It provides privacy for users and still allows the distance between users to be estimated. Distance is a useful parameter for many data mining algorithms. Balu et al. (2014) propose simple decoder and joint decoder attacks to reconstruct user profiles that are perturbed by BLIP and privacy-JLT methods.

4 Data reconstruction methods

In this section, we review data reconstruction methods that derive private data from perturbed data. There are different data reconstruction schemes proposed in the literature. We group major methods in the following subsections and present the related studies about each method.

4.1 Spectral theorem-based data reconstruction

Spectral filtering is one of the most widely used data reconstruction methods. Kargupta et al. (2003a, 2005) propose a random matrix-based spectral filtering technique that extracts original data from perturbed data using a random matrix theory for additive data perturbation. Because the randomization scheme requires knowing the type of distribution (uniform or Gaussian), the noise variance can be calculated from the given distribution. With the random matrix theory, the theoretical boundaries of the maximum/minimum values of the eigenvalues associated with noise matrix can be obtained. Meanwhile, the distribution of the eigenvalues of the sample covariance matrix calculated from the perturbed data can also be easily found. The authors claim that when the noise-related eigenvalues are extracted from the eigenvalues of the sample covariance matrix, the rest of the eigenvalues are used to retrieve the original data. The authors also note that the random perturbation techniques do not preserve privacy as much as desired. In many cases, the noise is easily extracted from the perturbed data using their method. In another study (Kargupta et al. 2003b), the authors study their previous work (Kargupta et al. 2003a) and improve on it by performing several experiments for discrete graph structured data.

Due to its popularity, Guo and Wu (2006a) and Guo et al. (2008) also conduct a theoretical study on spectral filtering and try to obtain an optimal performance to improve the spectral filtering method. They ask, "What bounds of the reconstruction error can be achieved by attackers using spectral filtering based techniques?" To answer this question, they propose an explicit upper bound for the reconstruction error using the matrix perturbation theory in Stewart and Sun (1990). The authors claim that the upper bound might be exploited by attackers using the spectral filtering method regardless of whether their estimate is close to the original data.

Principal component analysis (PCA) can be utilized to reduce the dimensions of a data set. Dimension reduction using PCA can be briefly explained as follows: (1) Given a data set, its covariance matrix is calculated. (2) The eigenvalues and the eigenvectors are obtained from

the covariance matrix. (3) The first p eigenvectors are chosen using the largest p eigenvalues, where $p < L$ and L is the number of variables used to represent data. Therefore, data can be represented using p variables rather than L variables in a reduced space. In terms of reconstruction, PCA can be utilized to exploit correlation among data items.

A PCA-based scheme is proposed as a data reconstruction method by Huang et al. (2005). The method is based on data correlations. The authors exploit the correlations among data to make more accurate guesses about the original data. PCA is directly utilized to reconstruct the data. Original eigenvalues and eigenvectors are first reconstructed. Then, principal components are selected based on the eigenvalues. Principal components and the sum of their variances indicate how much of the original data are kept via PCA. The PCA-based approach works well when data are highly correlated. If the correlations of data are not high enough, it is not easy to filter out a significant amount of noise. Thus, the scheme will not be able to reconstruct the original data.

Assuming that the attacker has a sample matrix (partial data from the original data as prior knowledge) and the mixing matrix is orthogonal, Liu et al. (2006b) also achieve a PCA-based attack. Their data reconstruction attack depends on finding the mixing matrix that comes from the same probability distribution function with the perturbed data. To test the distribution, the authors utilize the multivariate two-sample hypothesis proposed by Székely and Rizzo (2004). This is a two-sample test with null hypothesis $H_0 : F_1 = F_2$ and $H_1 : F_1 \neq F_2$. They target the rotation perturbation (Oliveira and Zaïane 2003b; Chen and Liu 2005) with their PCA-based attack.

Another PCA-based attack is proposed by Turgay et al. (2008). Their proposed data reconstruction method is on top of the hyper-lateration attack (discussed later in this section). It exploits PCA, whereby the attacker is equipped with the dissimilarity matrix and a representation of global distribution probability. A candidate data set is obtained from the dissimilarity matrix, randomly fixing $d + 1$ points (considering d-dimensional vector-space) by hyper-lateration. A PCA-based attack, similar to the one proposed by Liu et al. (2006b), is then applied to this candidate data set to match the target data set. The method is applicable to the random perturbation proposed by Liu et al. (2006a).

Singular value decomposition (SVD) divides a rectangular matrix A into a product of three matrices as $A_{n\times m} = U_{n\times n} S_{n\times m} V^T_{m\times m}$, where $U_{n\times n}$ and $V^T_{m\times m}$ are orthogonal matrices, $S_{n\times n}$ is a diagonal matrix with all singular values of the matrix A. The singular values can be calculated by taking the square roots of the eigenvalues of matrix $A^T A$ (or AA^T). All singular values are placed into the matrix S as a diagonal, and the rest of the matrix entries are filled with 0. Although $U_{n\times n}$ can be obtained from the eigenvectors of the matrix AA^T, $V^T_{m\times m}$ can be created by calculating the eigenvectors of the matrix $A^T A$.

An SVD-based data reconstruction method is proposed by Guo et al. (2006b, 2008). Their proposed method targets the additive data perturbation. The authors define a lower bound for the reconstruction error in addition to an upper bound for the reconstruction error. Given a privacy breach threshold, the reconstruction error bound is helpful for users and data miners to evaluate how much noise should be added. They also claim that the approach behind their SVD-based data reconstruction method is equal to the spectral filtering method. Briefly, SVD decomposes the perturbed data into three matrices like the spectral filtering method. One of the most important things in the SVD-based scheme is to determine the first k singular values of the perturbed data using $k = \min\left\{i \,|\, \tilde{\sigma}_i < \sqrt{2}\sigma_v\right\} - 1$, where $\tilde{\sigma}_i$ is the ith singular value of the perturbed data and σ_v is the largest singular value.

Zhang et al. (2006) also propose a data reconstruction method based on SVD to retrieve private information from randomly perturbed ratings. They develop their data reconstruction

method based on two random perturbation collaborative filtering (CF) schemes proposed by Polat and Du (2003, 2005). In their scheme, z-scores are used to normalize the rating matrix. The authors indicate that one of the powerful methods for describing the rating matrix is a low-dimensional linear model (Canny 2002). Hence, they declare that the sum of a low-dimensional linear model (X) and a Gaussian distributed noise matrix (Z) constitute the normalized matrix (Y_{norm}). R is generated using a Gaussian distribution. The perturbed data matrix can be represented as $Y = Y_{norm} + R = X + Z + R$. The authors argue that the log-likelihood of the z-scores of perturbed data given X, $logPr(Y|X)$, helps to find X because X maximizes the log-likelihood, $logPr(Y|X)$. The solution to finding X, which maximizes $logPr(Y|X)$, is a low rank approximation. If the rank of the linear model is k, the top k right singular vectors of Y can be produced by applying SVD.

4.2 ICA-based data reconstruction

ICA is a technique to observe linear representations of statistically independent components. Assume that the observed n random variables of $x = (x_1, x_2, ..., x_n)$ is a linear combination of m random variables of $s = (s_1, s_2, ..., s_m)$ as $x = As$ or $X = AS$, where A is a mixing matrix and s is called independent components (ICs), which is statistically independent and non-Gaussian by definition. These are the two main restrictions of ICA that are identifiable. n, which is the number of observed variables of x, and m, which is the number of random variables of s, are not necessarily same. However, for simplicity, the mixing matrix A can be considered square and invertible so that the number of observed variables of x and random variables of s could be equal and independent components can be obtained after Ahas been estimated. The mixing matrix of A can be estimated using an ICA model by exploiting the statistical properties of the perturbed variable of x. Such estimations should work for any distribution of ICs and some examples of these methods of ICA are the maximization of non-Gaussianity, maximum likelihood estimation, and minimization by mutual information (Hyvärinen et al. 2001).

ICA for multiplicative data perturbation is examined with respect to privacy breaches by Liu et al. (2006a). However, the authors argue that the ICA-based attack is not effective if the elements of the original data X are correlated and there is more than one Gaussian distribution. Chen and Liu (2005) make similar observations, stating that it is not suitable for data classification and ordering is not preserved. Similarly, Chen et al. (2007) study the ICA-based data reconstruction method. In addition to difficulties like correlation, Gaussian distribution, and ordering, they mention one more difficulty: Namely, ICA does not guarantee preserving the variance of the original data even if ordering is captured. Therefore, they claim that the attacker needs to have some basic statistics (mix/min values and the probability density function) to overcome the difficulties. They describe an enhanced ICA with these statistics to render an effective ICA-based data reconstruction attack.

According to ICA technique, where $X = AS = (A\Lambda P)(P^{-1}\Lambda^{-1}S)$ for any diagonal matrix Λ and permutation matrix P, the original data X can never have a unique representation (Liu 2007; Guo and Wu 2007). Guo and Wu (2007) propose a variation of an ICA-based attack, referred to as apriori-knowledge ICA (AK-ICA), for arbitrarily projected data, where the mixing matrix can be any matrix. They argue that S needs to have unit variance to avoid the scale issue of ICA. The AK-ICA attack reconstructs the original data by exploring relations on the sample and perturbed data by employing ICA for both of them. In their setting, the attacker is assumed to have a sample data matrix $\tilde{X}$ and perturbed data Y. In the AK-ICA approach, ICA representations, $\left(A_{\tilde{X}}, S_{\tilde{X}}\right)$ and (A_y, S_y) of the sample data $A_{\tilde{X}}$, and the perturbed data Y are obtained. A transformation matrix J is defined such that $\hat{X} = A_{\tilde{x}} J S_y$

is an estimate of the original data X. J is determined by analyzing distributions of S_x and S_Y by information difference measures to assess the similarity of two distributions. The AK-ICA method utilizes the relationship between perturbed and sample data through ICA. When compared to the PCA-based attack proposed by Liu et al. (2006b), their attack can be considered superior because the PCA-based attack requires an orthogonal transformation matrix, while the AK-ICA attack can be applied to any transformation matrix and is valid for different transformations. Guo and Wu (2006b) apply an AK-ICA-based data reconstruction approach for general linear transformation perturbed data without knowledge about noise.

A data reconstruction method based on undetermined independent component analysis (UICA) is investigated by Sang et al. (2009, 2012). The UICA data reconstruction method is utilized to recover data set X in cases where X is mixed with a rectangular mixing matrix; permutation and scaling ambiguities occur if the mixing matrix is not known (Liu 2007; Guo and Wu 2007). The authors target a multiplicative data perturbation (random projection) technique proposed by Liu et al. (2006a) and employ a UICA technique for data reconstruction when the mixing matrix is known with prior knowledge. The mixing matrix with prior knowledge, mean, and variance eliminates both permutation and scaling ambiguities. Sang et al. (2012) also discuss reconstructing original data from perturbed data in two different scenarios. These scenarios are based on the status of the data attributes of the original data and mixing matrix. In the first scenario, the data attributes are mutually independent and sparse, while the mixing matrix is not known and UICA is applied for the reconstruction. In the second case, the data attributes are not mutually dependent and not sparse with an unknown mixing matrix.

4.3 Distribution-based reconstruction

Earlier studies about data reconstruction focus on distribution reconstruction (Agrawal and Srikant 2000; Agrawal and Aggarwal 2001) for data mining purposes. Agrawal and Srikant (2000) propose a *Bayes posterior probability* approach and continue its iterative estimates until the difference between successive estimates becomes small enough. Agrawal and Aggarwal (2001) utilize the *expectation maximization* (EM) algorithm. They derive expectation and maximization algorithms that are specific to their implementation. Domingo-Ferrer et al. (2004) offer multivariate generalization of technique as an extension of the work presented by Agrawal and Srikant (2000). They propose an iterative algorithm computing the probabilities of original values that take values at certain intervals. They argue that the higher the dimensionality data is, the more likely the privacy disclosure is.

Guo et al. (2006a) introduce an *inter-quantile range* (IQR) attack on reconstructed distribution for additive data perturbation. The authors indicate that each individual piece of data is related to one privacy interval. Each data owner defines her privacy interval according to the privacy concerns, and each privacy interval may be different for each data owner's demand. Based on their approach, a random number should be large enough so that perturbed data will not be within the range of the user's privacy interval. On the other hand, reconstruction of the data distribution provides a certain level of knowledge; the attackers can derive sensitive data using the IQR. The IQR measures the spread of the data set; hence, an attacker can use IQR to estimate the range of each individual data. Their experiments include the distribution reconstruction of Agrawal and Srikant (2000) and spectral filtering with 95 % confidence for IQR. They state that they fully disclose the individual's data if the estimated range with IQR includes the private data and its privacy interval set by the owner of the data.

4.4 Bayesian statistics-based reconstruction

Bayes estimate-based data reconstruction (BE-DR) is proposed by Huang et al. (2005) as a data reconstruction technique. The authors formulate the data reconstruction problem as an estimation problem. Observed perturbed data and data correlations are utilized to reconstruct the original data. Given the perturbed data Y, the original data X is searched in such a way that the posterior distribution $P\ (X|Y)$ is maximized. Thus, the authors name their scheme BE-DR, which is actually based on the Bayes estimate. This reconstruction method is used for additive perturbation. Similarly, *maximum a posterior probability* (MAP) is first studied by Liu (2007), targeting the random projection (Liu et al. 2006a) to derive original data and quantify privacy. They are interested in estimating original data given perturbed data with MAP. Sang et al. (2009, 2012) also employ a MAP attack for random projection with a known and unknown mixing matrix when the data attributes are not mutually independent and not sparse.

4.5 Prior knowledge-based reconstruction

In this subsection, reconstruction techniques that rely on prior knowledge are listed if they do not exploit the techniques used in previous sections.

In case of known input-output pairs, Liu et al. (2006b) propose an attack technique where the attacker knows k columns of original data in addition to the perturbed data. The authors assume that the transformation matrix is orthogonal. Their method is based on choosing the right transformation matrix among possible candidates using ε-privacy breach probability $\rho\ (x, \varepsilon)$. They state an ε-privacy breach occurs if the distances between original and estimated data are more than the error ratio and define a formula for $\rho\ (x, \varepsilon)$. A probable attack chooses the best orthogonal matrix that will maximize privacy breach probability $\rho\ (x, \varepsilon)$. This attack is later extended by Giannella et al. (2013) with less but more realistic prior knowledge assumptions. They assume that the attacker has very few inputs (any number of inputs) and does not know their corresponding entries in the perturbed data. Their strategy is similar to the attack technique described above in choosing an orthogonal matrix and checking it against the privacy breach probability for the best one. However, a mapping between a known input and perturbed data has to be maintained beforehand, so the first step in this attack is to infer the original input data (prior knowledge) and perturbed counterparts. Both methods above are applied to the rotation perturbation proposed by Oliveira and Zaïane (2003b) and Chen and Liu (2005).

Calandrino et al. (2011) propose an attack to learn partial individual behavior in CF systems, where the attack is based on *passive inference* attacks. Their attack assumes that the attackers can access the recommender system's public outputs like item similarity lists, item-to-item co-variances, and relative popularity of items. To derive useful information about users, the attacker must observe the system over a certain period of time and capture changes in its outputs. The authors also assume that a subset of some users' transaction history is available to the attacker.

An optimal method exhausting all possibilities about an unknown probability to reconstruct trajectories is proposed by Kaplan et al. (2010). Their study for reconstructing trajectories is more related to geographic knowledge discovery. Giannella et al. (2013) compare their study to Kaplan et al. (2010). They show that an attacker with four known columns of original data and the corresponding columns in perturbed data can estimate an original unknown column with $<7\ \%$ error. Liu et al. (2008b) study whether a privacy breach can occur under some assumptions if the original records are known by the attacker. They emphasize

that angles between records, sub-rows, and sub-columns are preserved after general perturbation. If the original records are discovered by an attacker, this may yield unintended consequences for the perturbation method.

Hyper-lateration attacks are scrutinized by Turgay et al. (2008). Their hyper-lateration model is based on a secret data set of n objects drawn to a global probability of a $d$dimensional vector and a dissimilarity matrix P described by $P_{i,j} = ||X_i - X_j||$, where X_i and X_j are independent and identically distributed random variables. If the attacker has a prior knowledge about $d+1$ objects from the data set and their corresponding entries in the dissimilarity matrix, they can fully recover the original data set by utilizing the hyper-lateration method. The main idea behind hyper-lateration is trilateration. A fourth point with known distances to any given three points in a two-dimensional vector space (that do not lie on a line) can be placed by trilateration. Chen et al. (2007) discuss a similar approach with $d + 1$ prior knowledge for random data perturbation.

4.6 Others

Another data reconstruction method for additive data perturbation is proposed by Zhang et al. (2006), referred to as *k-means clustering*. The method retrieves private information from perturbed ratings. They divide their work into two cases based on whether the ratings are discrete- or continuous-valued. In the discrete-valued ratings, the authors first examine rated items only. They also investigate a case in which all entries are used. Before conducting k-means, unrated items should be defined and marked as unrated. They show how to determine the entries that could be unrated. When an item is unrated, its z-score is 0 in the normalized data. They claim that the distribution had already been known; hence, the unrated ratings whose z-scores are within the distribution range can be marked as unrated, and the remaining part is marked as rated. They also claim that some rated items may be marked as unrated or the opposite; however, precision is more prominent for attackers to evaluate their performance. The authors also apply k-means clustering when the ratings are continuous-valued.

Sramka et al. (2009) and Sramka (2010, 2012) illustrate an attack technique that is based on simple general-purpose data mining and fusion techniques using relational databases to retrieve sensitive information from published sanitized data. They claim that attackers do not have to know any background information and non-expert adversaries can easily perform this kind of attack to derive sensitive information. When the attacker can perform multiple data mining algorithms, different predictions should be obtained for the same individual data. Hence, it is essential to apply a fusion algorithm to obtain a single prediction from different miners. Two fusion algorithms are used in the proposed attack. The first one combines all distinct predictions for each miner, and the second one divides the sum of all predictions into the total number of data miners. The result of the second fusion algorithm gives an estimation for the original data.

Huang and Du (2008) discuss how to optimize privacy and utility provided by RRTs in PPDM. Utility describes how much of the data distribution can be reconstructed from perturbed data. The authors consider utility quantification as an estimation problem. However, instead of estimating individual data items, they estimate the distribution of the original data as aggregate information. Their goal is not to show how much aggregate data can be reconstructed. They actually utilize an optimization method to determine the optimal perturbing matrices for RRTs.

Balu et al. (2014) perform *single decoder* and *joint decoder attacks* targeting BLIP and privacy-JLT schemes. The main objective of these attacks is to reveal the presence or absence of particular items in a profile. The single decoder attack computes a score from the perturbed

data and checks the score against a threshold to test the presence or absence of the item in a user profile in a sequential way. Likewise, the joint decoder attack computes a score from the perturbed data for a profile, which is indeed a true profile. The difference between these two attacks is that the simple decoder attack starts with an item to constitute a reconstructed profile, while the joint decoder attack examines a perturbed profile that is a true profile.

Zhao et al. (2014) propose an l_1-*recovery* attack to reconstruct the original data ratings from the perturbed data by random projection. Their attack is on l_1-*minimization* when the random matrix is disclosed. They indicate that if the original data contains many zeros, it can be recovered exactly. If the original data is sparse, it can be approximated by an l_1-*recovery* attack.

Maximum entropy (ME) modeling is applied to derive confidential data from association mining results (Zhu et al. 2009). The authors investigate how much private information can be disclosed by association rule mining results if they perform an ME attack. The knowledge embedded in the data mining results is first converted into ME constraints. The proposed method is suitable for probability-based data mining results. Because association rule mining results are probability-based, the outcomes can be converted into constraints. Once the constraints are derived, they are then used to determine the probabilities of deriving sensitive attributes from given quasi-identifiers.

We might mention privacy analysis as another aspect of the effort being made to ensure privacy. In association rule mining, it has been extensively studied. Evfimievski et al. (2003) redefine the privacy breach made by Evfimievski et al. (2002). Evfimievski et al. (2003) introduce the notion of amplification, whereby obtaining the perturbed data does not reveal much about the private data in cases where obtaining a particular perturbed piece of data is almost equally likely for any private data. Rizvi and Haritsa (2002) study a scheme to assure privacy with accuracy, which is based on probabilistic distribution of the user data. Agrawal and Haritsa (2005) design a framework for perturbation schemes that support amplification (Evfimievski et al. 2003). Mukherjee et al. (2008) explore the work of Evfimievski et al. (2003), which is based on discrete data to continuous data. They also propose a combination of additive perturbation and PCA to protect distance-based algorithms against Liu et al. (2006b) and Huang et al. (2005) and privacy breaches as defined by Evfimievski et al. (2003).

5 Statistical analysis and evaluation

In this section, we first present a statistical overview of the research covered in this study to provide a broader perspective of the research direction. Then, we provide data sets and evaluation metrics in different subsections. Researchers investigating the data reconstruction methods of the data perturbation techniques discussed throughout this study demonstrate the effectiveness of their reconstruction in different ways. We believe that the data sets and metrics used to measure the successful evaluation of a reconstruction method will provide a better understanding of the proposed studies.

5.1 Statistical analysis

We present some statistical views of the studies we have covered. First, we start with the number of related publications over the years, as depicted in Fig. 1. In the year 2000, Agrawal and Srikant (2000) made the earliest effort in this area. Studies reach their highest level in 2006. Although we observe a decline in the number of studies after 2006, there have still been valuable contributions since then.

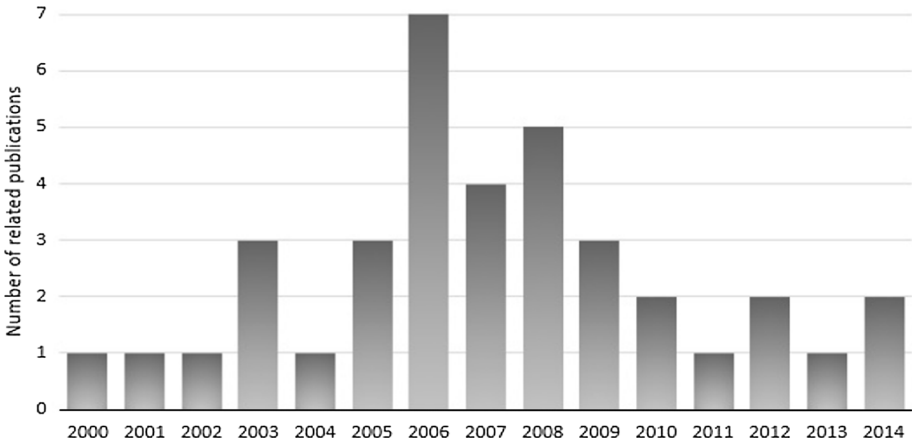


Fig. 1 Number of related publications over the years

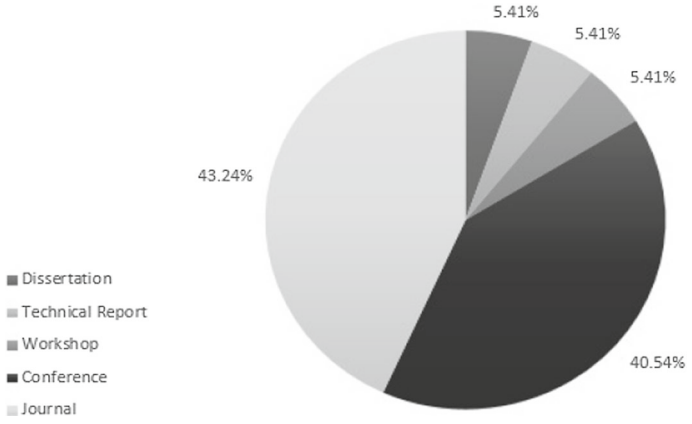


Fig. 2 Studies by publication type

Another group of statistics that we extract is the percentage of publications of the reconstruction methods, grouped by where they are published. As seen from Fig. 2, the majority of the publications, 43.24 %, are journal papers. Conference papers are approximately 41 %. Workshop papers, dissertations, and technical reports represent approximately 5.41 % of the publication percentage among all.

In Fig. 3, we also display the relevant studies based on the data reconstruction method that is used. The largest percentage is 26.83 % for spectral theorem-based methods. Other methods follow spectral theorem-based methods (19.51 %); however, it contains various methods in it. As seen from Fig. 3, ICA-based and prior knowledge-based approaches have been studied well with 17.07 %. Bayesian statistics and distribution-based methods are equally studied with 9.76 %.

We create a timeline of the studies by grouping them into the major methods used for reconstruction. Table 1 shows a timeline of the studies covered in this survey in chronological order by major methods used to demonstrate the evolution of the studies. We can conclude that the ICA-based attack has been popular since its emergence in 2006.

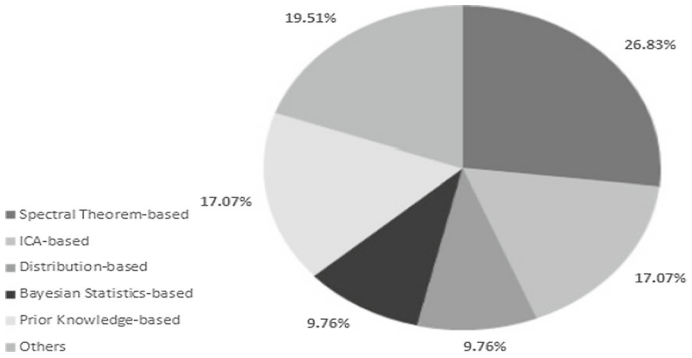


Fig. 3 Studies by reconstruction method

5.2 Data sets

While some researchers prefer to use well-known public data sets, others use synthetic data sets fitting the proposed problem. Table 2 shows the most commonly used data sets.

The Adult data set in the UCI repository, which is based on the census data (Bache and Lichman 2013), has been widely used in experiments by Guo et al. (2008), Liu et al. (2006b), Turgay et al. (2008), Sramka et al. (2009), and Sramka (2010, 2012). The set has 14 attributes and 32,561 instances; however, Guo et al. (2008) select six numerical attributes, and Liu et al. (2006b) select three continuous attributes. Sramka (2010, 2012) creates two modified Adult databases from the Adult data set, which are the Adult-UCI and Adult-Iyengar. The Adult-UCI database contains 45,222 records and 15 attributes. The Adult-Iyengar database is extracted from Adult-UCI by taking eight attributes with 30,162 records. The binary attribute salary class (a salary above or below 50,000) is also taken (Iyengar 2002). The Adult-Iyengar database is also used in Sramka et al. (2009). Records with missing values are discarded in both databases.

Sramka (2010, 2012) also creates two databases from the IPUMS97 data sets, which are IPUMS97-num and IPUMS97-cat. Both data sets, which have 61 attributes with 70,187 records, contain a sample of Los Angeles and Long Beach area citizens from 1997. While IPUMS97-num assumes all attributes are numerical, IPUMS97-cat assumes all attributes except age are categorical.

Another commonly used data set is Bank (Guo et al. 2006a; Guo and Wu 2006b, 2007), which was previously used in Muralidhar et al. (1999). The set consists of five attributes, which are Home Equity, Stock/Bonds, Liabilities, Savings, and CDs.

In addition to these data sets, there are other data sets like Auto-MPG, Letter Recognition, Ionosphere, Bupa Liver-disorders, Wine, Statlog, MovieLens, Jester, Digg, Acrene and DBworld. Turgay et al. (2008) use Auto-MPG (Bache and Lichman 2013) with 398 records and eight attributes. Giannella et al. (2013) use a Letter Recognition data set with 20,000 tuples and 16 numeric attributes (Bache and Lichman 2013). Kargupta et al. (2003a, 2005) use a real world single feature data set, which contains 10,000 sample points. Kargupta et al. (2003a) also conduct their experiments using Ionosphere data (Bache and Lichman 2013). Bupa Liver-disorders and Wine (Bache and Lichman 2013) data sets are used by Liu et al. (2008b). Bupa Liver-disorders contains five numerical and two categorical attributes with 345 instances. The Wine data set has 13 attributes in 178 instances. Sang et al. (2012) use Statlog (Bache and Lichman 2013) with seven attributes. Chen et al. (2007) use some UCI

Table 1 A timeline of studies

	2000	2001	2002	2003	2004	2005	2006	2007
Spectral theorem-based				Kargupta et al. (2003a, b)		Kargupta et al. (2005), Huang et al. (2005)	Guo and Wu (2006a); Liu et al. (2006b), Zhang et al. (2006), Guo et al. (2006b)	
ICA-based						Chen and Liu (2005)	Liu et al. (2006a); Guo and Wu (2006b)	Chen et al. (2007), Guo and Wu (2007)
Distribution-based	Agrawal and Srikant (2000)	Agrawal and Aggarwal (2001)			Domingo-Ferrer et al. (2004)		Guo et al. (2006a)	
Bayesian statistics-based						Huang et al. (2005)		Liu (2007)
Prior knowledge-based							Liu et al. (2006b)	Chen et al. (2007)
Others							Zhang et al. (2006)	

Table 1 continued

	2008	2009	2010	2011	2012	2013	2014
Spectral theorem-based	Guo et al. (2008), Turgay et al. (2008), Guo et al. (2008)						
ICA-based		Sang et al. (2009)			Sang et al. (2012)		
Distribution-based							
Bayesian statistics-based		Sang et al. (2009)			Sang et al. (2012)		
Prior knowledge-based	Liu et al. (2008b), Turgay et al. (2008)		Kaplan et al. (2010)	Calandrino et al. (2011)		Giannella et al. (2013)	
Others	Huang and Du (2008)	Zhu et al. (2009), Sramka et al. (2009)	Sramka (2010)		Sramka (2012)		Zhao et al. (2014), Balu et al. (2014)

Table 2 Data sets

Name	No. of attributes	Attribute characteristics	No. of instances
Adult	14	Categorical, integer	48,842
IPUMS	61	Categorical, integer	256,932
Bank	5	Integer	50,000

data sets. Zhang et al. (2006) use the well-known recommender systems data sets in their experiments, such as the MovieLens and Jester data sets. MovieLens is a movie recommendation website (www.grouplens.org), and each rating is based on a 5-star scale. They use the standard 100 K MovieLens data set including 943 users and 1862 items. Jester is an online joke recommender system. Jester data set consists of 3,000 users and 100 jokes. Each rating is from −10 to 10. (http://eigentaste.berkeley.edu/dataset/). Balu et al. (2014) test their attacks with 100K MovieLens and Digg data sets. The Digg data set includes users with the news they have read. It has 531 users and 1237 news items. Zhao et al. (2014) use Arcene and DBworld data sets with 10,000 real and 4702 binary attributes (Bache and Lichman 2013), respectively.

Although various real data sets are used in different works, synthetic data sets are also used to test the proposed methods. Agrawal and Srikant (2000) use a synthetic data generator (Agrawal et al. 1992) with 100,000 training records and 5000 test records. Agrawal and Aggarwal (2001) generate a data set with uniform and Gaussian distribution. Kargupta et al. (2003a, 2005) use an artificially generated data set, which contains 35 features and 300 instances. Guo et al. (2006a, b) use the same synthetic data set with Kargupta et al. (2003a). Guo and Wu (2006a) and Guo et al. (2008) also use the same artificial data; however, they only increase the size of instances to 30,000 instead of 300. Domingo-Ferrer et al. (2004) have an original data set with 5000 bivariate records. Huang et al. (2005) use different artificial data for their experiments. After computing the covariance matrix, the artificial data is formed using the covariance matrix. The size of the covariance matrix is different based on their different experiments. Liu et al. (2006b) utilize a synthetic data set containing 1,000 data points with mean (−10, 10) and covariance (1, 1.5; 1.5, 3). Turgay et al. (2008) apply a synthetic data set of 1000 with four to 12 attributes. Sang et al. (2009, 2012) use a synthetic data set with Laplace distribution. A 100,000 column synthetic data set generated from 100-variate Gaussian distribution is used by Giannella et al. (2013). A random data set with different sparsity parameters has been generated by Zhao et al. (2014) for experimental purposes.

5.3 Evaluation measures

Here, we briefly give several metrics that are used in the evaluation of the privacy measures. More detailed work was presented by Herlocker et al. (2004). In this section, evaluation metrics are grouped into several subcategories as Herlocker et al. (2004) did. These subcategories are predictive accuracy metrics, privacy disclosure and utility loss metrics, classification accuracy metrics and others. The first is the predictive accuracy metrics according to which we group the ones that measure how close the estimations are to the original data. The second subcategory contains the metrics related to privacy disclosure and utility loss. In the third subcategory, classification accuracy metrics are grouped that usually analyze the ratio if correct or incorrect predictions are made. The last category includes the rest of the metrics that we could not include in our three main categories.

5.3.1 Predictive accuracy metrics

Agrawal and Srikant (2000) propose a *privacy measure*. The aim of the quantification of privacy is to measure how closely the original distribution can be estimated. Their privacy measure is based on estimating a value in an interval with a confidence level of $c\%$. They use *classification accuracy* to evaluate the reconstructed models.

Average relative error between original and reconstructed data is defined by Guo and Wu (2006b, 2007), and similar definitions are made by different researchers (Sang et al. 2012; Liu et al. 2006b; Turgay et al. 2008) without explicitly stating the term *average relative error.* Similarly, *mean absolute error* is utilized by Zhang et al. (2006). They assess the reconstruction accuracy and show how close the reconstruction is to the original data. To evaluate the perturbation, *absolute error* and *relative error* in terms of the Frobenius norm are used (Guo et al. 2006a, b, 2008; Guo and Wu 2006a, b, 2007). *Relative error* is used to measure reconstruction errors as well as privacy loss.

Zhu et al. (2009) measure the closeness of estimated and original distributions at the individual and overall level. They use standard *Kullback–Lieber divergence* (Kullback and Leibler 1951) for both.

The success of the noise-removing attack is measured using two metrics by Sramka (2010, 2012). The first one is *cumulative added noise,* which sums the difference between the perturbed value and the original value for all distinct values. The second, *cumulative remaining noise*, sums the difference between the estimation value and the original data for all distinct values. The success ratio can be computed by dividing the cumulative remaining noise by the cumulative added noise.

Balu et al. (2014) exploit different metrics to evaluate their experiments. They use *cosine similarity* to measure the performance of the reconstructed profile against the original profile. Additionally, they quantify *mean Average Precision at R* (*mAP@R*) metric to assess the presence of an item. This metric calculates mean over the Q profile average prediction success. Zhao et al. (2014) use *privacy level*$-\varepsilon$ if original data can be approximated with high probability with $||x - \hat{x}||_2^2 \leq \varepsilon ||x||_2^2$, where x is the original data and $\hat{x}$ is the estimated data.

5.3.2 Privacy disclosure and utility loss metrics

Agrawal and Aggarwal (2001) show that the density function of the original data that can be revealed in the privacy measure proposed by Agrawal and Srikant (2000). Their proposed privacy measure focuses on privacy loss. They propose a metric based on the modification of *differential entropy* that gives the privacy of a random variable in terms of the length of a uniformly distributed random variable. They define *conditional privacy* using the differential entropy idea to identify privacy loss given when perturbed data is revealed. The *Information loss* metric is also used by Agrawal and Aggarwal (2001) to measure the information loss between the reconstructed and the original data. It is a universal metric in an interval of 0 and 1 that can be used in any data reconstruction algorithm.

The utility of data is evaluated using*the universal information loss* (Agrawal and Aggarwal 2001; Guo et al. 2008). The metric concentrates on the distribution similarity between the reconstructed and original data. The less the outcome is, the more successful the reconstruction is. Guo et al. (2006a) measure the disclosure ratio of individuals and compute *number of fully disclosed individuals* and *average disclosure* for evaluations.

Huang and Du (2008) quantify *privacy* and *utility*. *Privacy* is related to the individual information estimation of original data, while *utility* is related to the aggregate information

estimation of original data. *Privacy* is highly related to accuracy because higher accuracy means lower privacy. To quantify *utility*, average of *mean squared error (MSE)* is used. When original and estimated distributions are closer, *MSE* will result in smaller values, indicating better *utility.*

5.3.3 Classification accuracy metrics

Precision and *recall* are more related to the relevance of the estimations (Zhang et al. 2006). *Precision* depicts the ratio of the number of relevant estimations to the number of total estimations. *Recall* is the ratio of the number of relevant estimations to the total number of relevant entries.

The *Receiver operating characteristic-4* (*ROC-4)* area is used to assess prediction and accuracy (Zhang et al. 2006). *ROC* is a graphical plot that demonstrates the true positive rate and false positive rate. The *ROC-4* area is a type of problem specifically defined by the *ROC*, where the ratings that are less than four are considered noise.

Recovery rate with a threshold is used by Sang et al. (2009, 2012). According to their definition, it is a ratio defined by the number of reconstructed entries whose errors do not excess the threshold to the number of total entries in the original data.

5.3.4 Others

Signal-to-noise ratio (SNR) is the ratio of the variance of original data to the noise variance (Kargupta et al. 2003a, 2005). *SNR* is used to measure how much noise is added to the original data to perturb it. Kargupta et al. (2003b) also use modified *SNR* that is a special case for discrete graph structured data. *SNR* measures the ratio of the number of similar bits in the original data and the perturbed data to the number of dissimilar bits in the original data and the perturbed data.

Yield, number of inferences made, and *accuracy,* number of correct inferences, are used for evaluation by Calandrino et al. (2011). They show that *yield* can be evaluated as a proportion of inference.

Sramka (2010, 2012) measures the success of the estimation using a term called *nearness* (Sramka et al. 2009, 2010), which indicates whether the estimation is better for the original value than the perturbed value. If an estimate made for a perturbed value is closer to the original value than the perturbed value, the estimate is called a nearer estimate. Otherwise, it is not near.

6 Discussion—open questions

There are two main classes of perturbation techniques, namely, additive and multiplicative perturbations. These classes have different variations, from randomization and rotation perturbation to random projection. The perturbation techniques offered by these perturbation methods primarily address numeric data. Various reconstruction techniques that can extract noise such as spectral filtering (Kargupta et al. 2003a, b, 2005), PCA (Huang et al. 2005), and SVD (Guo et al. 2006b, 2008) have been shown to be successful against randomization. Maximum/minimum bounds of eigenvalues of noise can be obtained by spectral filtering-based approaches, while PCA works well if data is highly correlated and SVD can define lower and upper bounds for reconstruction error. Metrics such as SNR and privacy breach have been

proposed to evaluate the amount of noise added while determining reconstruction success. Similarly, the data owner should ponder how much noise must be added while performing additive perturbation to meet both privacy and accuracy goals. On the other hand, multiplicative perturbation techniques are also subjected to some different reconstruction methods in the literature mainly centered on the ICA model (Liu et al. 2006a; Chen and Liu 2005; Chen et al. 2007; Guo and Wu 2007). However, the ICA model is not considered effective due to the limitations discussed by scholars and the amount of prior knowledge to avoid limitations. Reconstruction techniques based on prior knowledge of the known input-output relation are also studied (Liu et al. 2006b; Giannella et al. 2013). Known input-output-based attacks are effective even if very little input is known (Giannella et al. 2013).

Researchers have been presenting various attacks to derive information from perturbed data, and this trend is expected to grow in parallel with the increasing attention placed on PPDM. Although these studies reveal vulnerabilities in perturbation methods, there are still some issues associated with these attacks and the resilience of perturbation methods. We will list some open questions that we believe researchers need to address in future studies to consolidate more robust perturbation methods as follows:

1. Researchers mainly focus on deriving numerical ratings from perturbed data; however, deriving binary ratings has not received much attention. Huang and Du (2008) perform a work that optimizes privacy and utility for RRTs. Therefore, vulnerabilities of binary ratings and deriving private information from the perturbed data that contain binary ratings need to be further examined. Special attention should be given to deriving binary ratings from perturbed data.
2. Prior knowledge such as the known input-output and known sample are generally used for data reconstruction; however, we believe that other prior knowledge can also be considered informative while deriving information from perturbed data. For example, extreme data values and auxiliary information obtained from other sources can disclose important information. The range of prior knowledge that can be exploited in the reconstruction of private data must be investigated.
3. Although some researchers achieve good results to reconstruct original data, the question is how reliable their assumptions are. Most of the studies rely on prior knowledge of the known input-output and known sample. If the data owner somehow realizes the prior knowledge that adversaries have, she can easily neglect it to eliminate privacy risks. In brief, schemes relying on well-founded assumptions should be revisited.
4. Although Evfimievski et al. (2003), Rizvi and Haritsa (2002), Agrawal and Haritsa (2005), and Mukherjee et al. (2008) study privacy quantification, more attention should be given to privacy analysis. Privacy analysis is important to derive a scheme to assess the resilience of perturbation methods. Distance-preserving algorithms in particular lack such analysis. We believe analysis evaluating current perturbation methods is imperative.

7 Conclusions and future work

In this work, we place special attention on those studies focusing on the data reconstruction methods proposed to derive confidential data from perturbed data. First, we briefly explain the related survey papers addressing data reconstruction schemes and discuss our differences with those studies. Unlike such surveys, our review covers the works that span from the beginning to today. We also discuss those data perturbation methods that are subjected to data reconstruction attacks. Second, we study the data perturbation methods that are attacked by

data reconstruction techniques. Notice that our aim is to briefly describe such schemes rather than explain them in detail. Third, we investigate data reconstruction methods. We group well-known and widely used schemes into major classes and explain the others separately. Related studies about the data reconstruction methods are briefly listed. Fourth, we present statistics, evaluation metrics, and the data sets used to measure the proposed models. Basic statistics gives an overall picture of the researchers and their studies in the field. Additionally, the timeline shows the milestone studies in the literature. Evaluation metrics and the data sets help researchers easily conduct their studies and compare them with existing ones. We finally include some open problems for researchers who want to devote more attention to the field.

Our survey provides an analysis and review of how to derive private information from perturbed data. We believe that this survey serves us and other researchers as a necessary reference for further study. Our intention is to look for other vulnerabilities in perturbation methods that can be exploited so that we can provide some suggestions to improve privacy levels by modifying the related perturbation methods. In addition, enhanced data perturbation methods and related attacking schemes might be proposed.

Acknowledgments This work is supported by Grant 113E262 from TUBITAK.

References

Aggarwal CC, Yu PS (2008) A survey of randomization methods for privacy preserving data mining. In: Aggarwal CC, Yu PS (eds) Privacy-preserving data mining: models and algorithms. Springer, New York, pp 137–156

Agrawal D, Aggarwal CC (2001) On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems. Santa Barbara, pp 247–255

Agrawal R, Ghosh SP, Imielinski T, Iyer BR, Swami A (1992) An interval classifier for database mining applications. In: Proceedings of the 18th international conference on very large databases. Vancouver, British Columbia, pp 560–573

Agrawal R, Srikant R (2000) Privacy-preserving data mining. In: Proceedings of the 2000 ACM SIGMOD international conference on management of data. Dallas, pp 439–450

Agrawal S, Haritsa JR (2005) A framework for high-accuracy privacy-preserving mining. In: Proceedings of 21st international conference on data engineering. Los Alamitos, pp 193–204

Alaggan M, Gambs S, Kermarrec A-M (2012) BLIP: non-interactive differentially-private similarity computation on bloom filters. Lecture notes in computer science, vol 7596. pp 202–216. doi:10.1007/978-3-642-33536-5_20

Amiri A (2007) Dare to share: protecting sensitive knowledge with data sanitization. Decis Support Syst 43(1):181–191. doi:10.1016/j.dss.2006.08.007

Atallah M, Elmagarmid A, Ibrahim M, Bertino E, Verykios V (1999) Disclosure limitation of sensitive rules. In: Proceedings of the 1999 workshop on knowledge and data engineering exchange. Chicago, pp 45–52

Bache K, Lichman M (2013) UCI machine learning repository. http://archive.ics.uci.edu/ml. Accessed 22 Sept 2013

Balu R, Furon T, Gambs S (2014) Challenging differential privacy: The case of non-interactive mechanisms. Lecture notes in computer science, vol 8713. pp 146–164. doi:10.1007/978-3-319-11212-1_9

Calandrino JA, Kilzer A, Narayanan A, Felten EW, Shmatikov V (2011) You might also like: privacy risks of collaborative filtering. In: Proceedings of the 2011 IEEE symposium on security and privacy. Berkeley, pp 231–246

Canny J (2002) Collaborative filtering with privacy via factor analysis. In: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval. Tampere, pp 238–245

Chen K, Liu L (2005) Privacy preserving data classification with rotation perturbation. In: Proceedings of the 5th IEEE international conference on data mining. Houston, pp 589–592

Chen K, Sun G, Liu L (2007) Towards attack-resilient geometric data perturbation. In: Proceedings of the 2007 SIAM international conference on data mining. Minneapolis, pp 78–89

Chen K, Liu L (2008) A survey of multiplicative perturbation for privacy preserving data mining. In: Aggarwal CC, Yu PS (eds) Privacy-preserving data mining: models and algorithms. Springer, New York, pp 157–181

Domingo-Ferrer J, Sebé F, Castellà-Roca J (2004) On the security of noise addition for privacy in statistical databases. Lecture notes in computer science, vol 3050. pp 149–161. doi:10.1007/978-3-540-25955-8_12

Du W, Zhan Z (2003) Using randomized response techniques for privacy-preserving data mining. In: Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining. Washington, pp 505–510

Evfimievski A, Srikant R, Agrawal R, Gehrke J (2002) Privacy preserving mining of association rules. In: Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining. Edmonton, pp 217–228

Evfimievski A, Gehrke J, Srikant R (2003) Limiting privacy breaches in privacy preserving data mining. In: Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems. San Diego, pp 211–222

Giannella CR, Liu K, Kargupta H (2013) Breaching Euclidean distance-preserving data perturbation using few known inputs. Data Knowl Eng 83:93–110. doi:10.1016/j.datak.2012.10.004

Guo L, Wu X (2009) Privacy preserving categorical data analysis with unknown distortion parameters. Trans Data Priv 2:185–205

Guo S, Wu X (2006a) On the use of spectral filtering for privacy preserving data mining. In: Proceedings of the 21st annual ACM symposium on applied computing. Dijon, pp 622–626

Guo S, Wu X (2006b) Deriving private information from general linear transformation perturbed data. Technical report, The University of North Carolina at Charlotte, Charlotte

Guo S, Wu X, Li Y (2006a) Deriving private information from perturbed data using IQR based approach. In: Proceedings of the 22nd international conference on data engineering workshops. Atlanta, pp 92–101

Guo S, Wu X, Li Y (2006b) On the lower bound of reconstruction error for spectral filtering based privacy preserving data mining. Lecture notes in computer science, vol 4213. pp 520–527. doi:10.1007/11871637_51

Guo S (2007) Analysis of and techniques for privacy preserving data mining. Dissertation, University of North Carolina at Charlotte

Guo S, Wu X (2007) Deriving private information from arbitrarily projected data. Lecture notes in computer science, vol 4426. pp 84–95. doi:10.1007/978-3-540-71701-0_11

Guo S, Wu X, Li Y (2008) Determining error bounds for spectral filtering based reconstruction methods in privacy preserving data mining. Knowl Inf Syst 17(2):217–240. doi:10.1007/s10115-008-0123-9

Herlocker JL, Konstan JA, Terveen LG, Riedl JT (2004) Evaluating collaborative filtering recommender systems. ACM Trans Inform Syst 22(1):5–53. doi:10.1145/963770.963772

Huang Z, Du W, Chen B (2005) Deriving private information from randomized data. In: Proceedings of the 2005 ACM SIGMOD international conference on management of data. Baltimore, pp 37–48

Huang Z, Du W (2008) OptRR: optimizing randomized response schemes for privacy-preserving data mining. In: Proceedings of the 2008 IEEE 24th international conference on data engineering. Cancun, pp 705–714

Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley, New York

Iyengar VS (2002) Transforming data to satisfy privacy constraints. In: Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining. Edmonton, pp 279–288

Johnson WB, Lindenstrauss J (1984) Extension of Lipshitz mappings into Hilbert space. Contemp Math 26:189–206. doi:10.1090/conm/026/737400

Kaplan E, Pedersen TB, Savas E, Saygin Y (2010) Discovering private trajectories using background information. Data Knowl Eng 69(7):723–736. doi:10.1016/j.datak.2010.02.008

Kargupta H, Datta S, Wang Q, Sivakumar K (2003a) On the privacy preserving properties of random data perturbation techniques. In: Proceedings of the 3rd IEEE international conference on data mining. Melbourne, pp 99–106

Kargupta H, Dutta H, Datta S, Sivakumar K (2003) Analysis of privacy preserving random perturbation techniques: further explorations. In: Proceedings of the 2003 ACM workshop on privacy in the electronic society. Washington, pp 31–38

Kargupta H, Datta S, Wang Q, Sivakumar K (2005) Random-data perturbation techniques and privacy-preserving data mining. Knowl Inf Syst 7(4):387–414. doi:10.1007/s10115-004-0173-6

Kenthapadi K, Korolova A, Mironov I, Mishra N (2013) Privacy via the Johnson–Lindenstrauss transform. J Priv Confid 5(1):39–71

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86. doi:10.1214/aoms/1177729694
Liu K, Kargupta H, Ryan J (2006) Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Tran Knowl Data Eng 18(1):92–106. doi:10.1109/TKDE.2006.14
Liu K, Giannella C, Kargupta H (2006b) An attacker's view of distance preserving maps for privacy preserving data mining. Lecture notes in computer science, vol 4213. pp 297–308. doi:10.1007/11871637_30
Liu K (2007) Multiplicative data perturbation for privacy preserving data mining. Dissertation, University of Maryland, Baltimore County
Liu K, Giannella C, Kargupta H (2008a) A survey of attack techniques on privacy-preserving data perturbation methods. In: Aggarwal CC, Yu PS (eds) Privacy-preserving data mining: models and algorithms. Springer, New York, pp 359–381
Liu L, Wang J, Zhang J (2008b) Privacy vulnerabilities with background information in data perturbation. Technical report. Department of Computer Science, University of Kentucky
Mukherjee S, Banerjee S, Chen Z, Gangopadhyay A (2008) A privacy preserving technique for distance-based classification with worst case privacy guarantees. Data Knowl Eng 66(2):264–268. doi:10.1016/j.datak.2008.03.004
Muralidhar K, Parsa R, Sarathy R (1999) A general additive data perturbation method for database security. Manage Sci 45(10):1399–1415. doi:10.1287/mnsc.45.10.1399
Oliveira SRM, Zaïane OR (2002) Privacy preserving frequent itemset mining. In: Proceedings of the IEEE international conference on privacy. Security and data mining. Maebashi City, pp 43–54
Oliveira SRM, Zaïane OR (2003a) Protecting sensitive knowledge by data sanitization. In: Proceedings of the 3rd IEEE international conference on data mining. Melbourne, pp 613–616
Oliveira SRM, Zaïane OR (2003b) Privacy preserving clustering by data transformation. In: Proceedings of the 18th Brazilian symposium on databases. Manaus, pp 304–318
Polat H, Du W (2003) Privacy-preserving collaborative filtering using randomized perturbation techniques. In: Proceedings of the 3rd IEEE international conference on data mining. Melbourne, pp 625–628
Polat H, Du W (2005) SVD-based collaborative filtering with privacy. In: Proceedings of the 21st annual ACM symposium on applied computing. Dijon, pp 791–795
Polat H, Du W (2006) Achieving private recommendations using randomized response techniques. Lecture notes in computer science, vol 3918. pp 637–646. doi:10.1007/11731139_73
Rizvi SJ, Haritsa JR (2002) Maintaining data privacy in association rule mining. Proceedings of the 28th international conference on very large data bases. Hong Kong, pp 682–693
Sang Y, Shen H, Tian H (2009) Reconstructing data perturbed by random projections when the mixing matrix is known. Lecture notes in computer science, vol 5782. pp 334–349. doi:10.1007/978-3-642-04174-7_22
Sang Y, Shen H, Tian H (2012) Effective reconstruction of data perturbed by random projections. IEEE Trans Comput 61(1):101–117. doi:10.1109/TC.2011.83
Saygin Y, Verykios VS, Clifton C (2001) Using unknowns to prevent discovery of association rules. SIGMOD Rec 30(4):45–54. doi:10.1145/604264.604271
Sramka M, Safavi-Naini R, Denzinger J (2009) An attack on the privacy of sanitized data that fuses the outputs of multiple data miners. In: Proceedings of the 9th IEEE international conference on data mining workshops. Miami, pp 130–137
Sramka M (2010) A privacy attack that removes the majority of the noise from perturbed data. In: Proceedings of the 2010 international joint conference on neural networks. Barcelona, pp 1–8
Sramka M, Safavi-Naini R, Denzinger J, Askari M (2010) A practice-oriented framework for measuring privacy and utility in data sanitization systems. In: Proceedings of the 12th international conference on extending database technology workshops. Lausanne
Sramka M (2012) Breaching privacy using data mining: removing noise from perturbed data. Stud Comput Intell 394:135–157. doi:10.1007/978-3-642-25237-2_9
Stewart GW, Sun J (1990) Matrix perturbation theory. Academic Press, Waltham
Székely GJ, Rizzo ML (2004) Testing for equal distributions in high dimension. InterStat 5:1–6
Turgay EO, Pedersen TB, Saygin Y, Savas E, Levi A (2008) Disclosure risks of distance preserving data transformations. Lecture notes in computer science, vol 5069. pp 79–94. doi:10.1007/978-3-540-69497-7_8
Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc 60(309):63–69. doi:10.1080/01621459.1965.10480775
Zhang S, Ford J, Makedon F (2006) Deriving private information from randomly perturbed ratings. In: Proceedings of the 6th SIAM international conference on data mining. Bethesda, pp 59–69

Zhao J, Yang J, Zhang J (2014) Privacy properties of random projection perturbation when random matrix is leaking. J Comput Inf Syst 10(8):3465–3472
Zhu Z, Wang G, Du W (2009) Deriving private information from association rule mining results. In: Proceedings of the 25th international conference on data engineering. Shanghai, pp 18–29