Data Min Knowl Disc
DOI 10.1007/s10618-008-0119-9

FRAPP: a framework for high-accuracy privacy-preserving mining

Shipra Agrawal · Jayant R. Haritsa · B. Aditya Prakash

Received: 16 July 2006 / Accepted: 26 September 2008
Springer Science+Business Media, LLC 2008

Abstract To preserve client privacy in the data mining process, a variety of techniques based on random perturbation of individual data records have been proposed recently. In this paper, we present FRAPP, a generalized matrix-theoretic framework of random perturbation, which facilitates a systematic approach to the design of perturbation mechanisms for privacy-preserving mining. Specifically, FRAPP is used to demonstrate that (a) the prior techniques differ only in their choices for the perturbation matrix elements, and (b) a symmetric positive-definite perturbation matrix with minimal condition number can be identified, substantially enhancing the accuracy even under strict privacy requirements. We also propose a novel perturbation mechanism wherein the matrix elements are themselves characterized as random variables, and demonstrate that this feature provides significant improvements in privacy at only a marginal reduction in accuracy. The quantitative utility of FRAPP, which is a general-purpose random-perturbation-based privacy-preserving mining technique, is evaluated specifically with regard to association and classification rule mining on

Responsible editor: Johannes Gehrke.

A partial and preliminary version of this paper appeared in the Proc. of the 21st IEEE Intl. Conf. on Data Engineering (ICDE), Tokyo, Japan, 2005, pgs. 193–204.

S. Agrawal · J. R. Haritsa (✉)
Indian Institute of Science, Bangalore 560012, India
e-mail: haritsa@dsl.serc.iisc.ernet.in

Present Address:
S. Agrawal
Stanford University, Stanford, CA, USA

B. A. Prakash
Indian Institute of Technology, Mumbai 400076, India

a variety of real datasets. Our experimental results indicate that, for a given privacy requirement, either substantially lower modeling errors are incurred as compared to the prior techniques, or the errors are comparable to those of direct mining on the true database.

Keywords Privacy · Data mining

1 Introduction

The knowledge models produced through data mining techniques are only as good as the accuracy of their input data. One source of data inaccuracy is when users, due to privacy concerns, deliberately provide wrong information. This is especially common with regard to customers asked to provide personal information on Web forms to E-commerce service providers. The standard approach to address this problem is for the service providers to assure the users that the databases obtained from their information would be anonymized through the variety of techniques proposed in the statistical database literature (see Adam and Wortman 1989; Shoshani 1982), before being supplied to the data miners. For example, the swapping of values between different customer records, as proposed by Denning (1982). However, in today's world, most users are (perhaps justifiably) cynical about such assurances, and it is therefore imperative to demonstrably provide privacy at the point of data collection itself, that is, *at the user site*.

For the above "B2C (business-to-customer)" privacy environment (Zhang et al. 2004), a variety of privacy-preserving data mining techniques have been proposed in the last few years (e.g. Aggarwal and Yu 2004; Agrawal and Srikant 2000; Evfimievski et al. 2002; Rizvi and Haritsa 2002), in an effort to encourage users to submit correct inputs. The goal of these techniques is to ensure the privacy of the raw local data but, at the same time, support accurate reconstruction of the global data mining models. Most of the techniques are based on a *data perturbation* approach, wherein the user data is distorted in a probabilistic manner that is disclosed to the eventual miner. For example, in the MASK technique Rizvi and Haritsa (2002), intended for privacy-preserving association-rule mining on sparse boolean databases, each bit in the original (true) user transaction vector is independently flipped with a parametrized probability.

1.1 The FRAPP framework

The trend in the prior literature has been to propose *specific* perturbation techniques, which are then analyzed for their privacy and accuracy properties. We move on, in this paper, to proposing FRAPP[1] (FRamework for Accuracy in Privacy-Preserving mining), a generalized matrix-theoretic *framework* that facilitates a systematic approach to the *design* of random perturbation schemes for privacy-preserving mining. It supports "amplification", a particularly strong notion of privacy proposed by

[1] Also the name of a popular coffee-based beverage, where the ingredients are perturbed and hidden under foam http://en.wikibooks.org/wiki/Cookbook:Frapp%C3%A9_Coffee.

Evfimievski et al. (2003), which guarantees strict limits on privacy breaches of individual user information, *independent of the distribution of the original* (*true*) *data*. The distinguishing feature of FRAPP is its quantitative characterization of the *sources of error* in the random data perturbation and model reconstruction processes.

We first demonstrate that the prior techniques differ only in their choices for the elements in the FRAPP perturbation matrix. Next, and more importantly, we show that through appropriate choices of matrix elements, new perturbation techniques can be constructed that provide highly accurate mining results even under strict amplification-based (Evfimievski et al. 2003) privacy guarantees. In fact, we identify a perturbation matrix with provably *minimal condition number*,[2] substantially improving the accuracy under the given constraints. An efficient implementation for this optimal perturbation matrix is also presented.

FRAPP's quantification of reconstruction error highlights that, apart from the choice of perturbation matrix, the size of the dataset also has significant impact on the accuracy of the mining model. We explicitly characterize this relationship, thus aiding the miner decide the minimum amount of data to be collected in order to achieve, with high probability, a desired level of accuracy in the mining results. Further, for those environments where data collection possibilities are limited, we propose a novel "multi-distortion" method that makes up for the lack of data by collecting multiple distorted versions from each individual user without materially compromising on privacy.

We then investigate, for the first time, the possibility of *randomizing the perturbation parameters themselves*. The motivation is that it could result in increased privacy levels since the actual parameter values used by a specific client will not be known to the data miner. This approach has the obvious downside of perhaps reducing the model reconstruction accuracy. However, our investigation shows that the trade-off is very attractive in that the privacy increase is significant whereas the accuracy reduction is only marginal. This opens up the possibility of using FRAPP in a *two-step* process: First, given a user-desired level of privacy, identifying the deterministic values of the FRAPP parameters that both guarantee this privacy and also maximize the accuracy; and then, (optionally) randomizing these parameters to obtain even better privacy guarantees at a minimal cost in accuracy.

1.2 Evaluation of FRAPP

The FRAPP model is valid for random-perturbation-based privacy-preserving mining in general. Here, we focus on its applications to *categorical databases*, where attribute domains are finite. Note that boolean data is a special case of this class, and further, that continuous-valued attributes can be converted into categorical attributes by partitioning the domain of the attribute into fixed length intervals. To quantitatively assess FRAPP's utility, we specifically evaluate the performance of our new perturbation mechanisms on popular mining tasks such as *association rule mining* and *classification rule mining*.

[2] In the class of symmetric positive-definite matrices (refer Sect. 4).

With regard to association rule mining, our experiments on a variety of real datasets indicate that FRAPP is substantially more accurate than the prior privacy-preserving techniques. Further, while their accuracy degrades with increasing itemset length, FRAPP is almost *impervious* to this parameter, making it particularly well-suited to datasets where the lengths of the maximal frequent itemsets are comparable to the cardinality of the set of attributes requiring privacy. Similarly, with regard to classification rule mining, our experiments show that FRAPP provides an accuracy that is, in fact, comparable to *direct classification on the true database*.

Apart from mining accuracy, the *running time* and *memory* costs for perturbed data mining, as compared to classical mining on the original data, are also important considerations. In contrast to much of the earlier literature, FRAPP uses a generalized *dependent* perturbation scheme, where the perturbation of an attribute value may be affected by the perturbations of the other attributes in the same record. However, we show that it is fully decomposable into the perturbation of individual attributes, and hence has the *same run-time complexity* as any independent perturbation method. Further, we present experimental evidence that FRAPP takes only a few minutes to perturb datasets running to millions of records. Subsequently, due to its well-conditioned and trivially invertible perturbation matrix, FRAPP incurs only *negligible* additional overheads with respect to memory usage and mining execution time, as compared to traditional mining. Overall, therefore, FRAPP does not pose any significant additional computational burdens on the data mining process.

1.3 Contributions

In a nutshell, the work presented here provides mathematical and algorithmic foundations for efficiently providing both strict privacy and enhanced accuracy in privacy-conscious data mining applications. Specifically, our main contributions are as follows:

– FRAPP, a generalized matrix-theoretic framework for random perturbation and mining model reconstruction;
– Using FRAPP to derive new perturbation mechanisms for minimizing the model reconstruction error while ensuring strict privacy guarantees;
– Introducing the concept of randomization of perturbation parameters, and thereby deriving enhanced privacy;
– Efficient implementations of the proposed perturbation mechanisms;
– Quantitatively demonstrating the utility of FRAPP in the context of association and classification rule mining.

1.4 Organization

The remainder of this paper is organized as follows: Related work on privacy-preserving mining is reviewed in Sect. 2. The FRAPP framework for data perturbation and model reconstruction is presented in Sect. 3. Appropriate choices of the framework parameters for simultaneously guaranteeing strict data privacy and improving model accuracy are discussed in Sects. 4 and 5. The impact of randomizing the FRAPP parameters is investigated in Sect. 6.

Efficient schemes for implementing the FRAPP approach are described in Sect. 7. The application of these mechanisms to specific patterns is discussed in Sect. 8, and their utility is quantitatively evaluated in Sect. 9. Finally, in Sect. 10, we summarize the conclusions of our study and outline future research avenues.

2 Related work

The issue of maintaining privacy in data mining has attracted considerable attention over the last few years. The literature closest to our approach includes that of Agrawal and Aggarwal (2001), Agrawal and Srikant (2000), de Wolf et al. (1998), Evfimievski et al. (2002, 2003), Kargupta et al. (2003), Rizvi and Haritsa (2002). In the pioneering work of Agrawal and Srikant (2000), privacy-preserving data classifiers based on adding noise to the record values were proposed. This approach was extended by Agrawal and Aggarwal (2001) and Kargupta et al. (2003) to address a variety of subtle privacy loopholes.

New randomization operators for maintaining data privacy for boolean data were presented and analyzed by Evfimievski et al. (2002), Rizvi and Haritsa (2002). These methods are applicable to categorical/boolean data and are based on probabilistic mapping from the domain space to the range space, rather than by incorporating additive noise to continuous-valued data. A theoretical formulation of privacy breaches for such methods, and a methodology for limiting them, were given in the foundational work of Evfimievski et al. (2003).

Techniques for data hiding using perturbation matrices have also been investigated in the statistics literature. For example, in the early 90s work of Duncan and Pearson (1991), various disclosure-limitation methods for microdata are formulated as “matrix masking” methods. Here, the data consumer is provided the masked data file $M = AXB + C$ instead of the true data X, with A, B and C being masking matrices. But, no quantification of privacy guarantees or reconstruction errors was discussed in their analysis.

The PRAM method (de Wolf et al. 1998; Gouweleeuw et al. 1998), also intended for disclosure limitation in microdata files, considers the use of Markovian perturbation matrices. However, the ideal choice of matrix is left as an open research issue, and an iterative refinement process to produce acceptable matrices is proposed as an alternative. They also discuss the possibility of developing perturbation matrices such that data mining can be carried out *directly on the perturbed database* (that is, as if it were the original database and therefore not requiring any matrix inversion), and still produce accurate results. While this “invariant PRAM”, as they call it, is certainly an attractive notion, the systematic identification of such matrices and the conditions on their applicability is still an open research issue—moreover, it appears to be feasible only in a “B2B (business-to-business)” environment, as opposed to the B2C environment considered here.

The work recently presented by Agrawal et al. (2005) for ensuring privacy in the OLAP environment, also models data perturbation and reconstruction as matrix-theoretic operations. A transition matrix is used for perturbation, and reconstruction is executed using matrix inversion. They also suggest that the condition number of the perturbation matrix is a good indicator of the error in reconstruction. However the issue of choosing a perturbation matrix to minimize this error is not addressed.

Our work extends the above-mentioned methodologies for privacy-preserving mining in a variety of ways. First, we combine the various approaches for random perturbation on categorical data into a common theoretical framework, and explore how well random perturbation methods can perform in the face of strict privacy requirements. Second, through quantification of privacy and accuracy measures, we present an ideal choice of perturbation matrix, thereby taking the PRAM approach to, in a sense, its logical conclusion. Third, we propose the idea of randomizing the perturbation matrix elements themselves, which has not been, to the best of our knowledge, previously discussed in the literature.

Very recently, Rastogi et al. (2007) utilize and extend the FRAPP framework to a B2B environment like publishing. That is, they assume that users provide correct data to a central server and then this data is collectively anonymized. In contrast, our schemes assume that the users trust no one but themselves, and therefore the perturbation has to happen locally for each user. Formally, the transformation in their algorithm is described as $y = Ax+b$, thereby effectively adding a noise vector b to Ax. They also analyze the privacy and accuracy tradeoff under bounded prior knowledge assumptions.

The "sketching" methods that were very recently presented by Mishra and Sandler (2006) are complementary to our approach. Their basic idea is that a k-bit attribute with 2^k possible values can be represented using 2^k binary-valued attributes which can then each be perturbed independently. However, a direct application of this idea requires extra $(2^k - k)$ bits, and therefore, Mishra and Sandler (2006) proposes a summary sketching technique that requires an extra number of bits logarithmic in the number of instances in the dataset. Due to the extra bits, the method provides good estimation accuracy for *single* item counts. However, the multiple-attribute count estimation accuracy is shown to depend on the condition number of the perturbation matrix. Our results on optimally conditioned perturbation matrices can be combined with the sketching methods to provide better estimation of joint distributions. Another difference between the two works is that we provide experimental results in addition to the theoretical formulations.

Recently, a new privacy-preserving scheme based on the interesting idea of *algebraic distortion*, rather than statistical methods, has been proposed by Zhang et al. (2004). Their work is based on the assumption that statistical methods cannot handle long frequent itemsets. But, as shown in this paper, FRAPP successfully finds even length-7 frequent itemsets. A second assumption is that each attribute is randomized independently, thereby losing correlations—however, FRAPP supports dependent attribute perturbation and can therefore preserve correlations quite effectively. Finally, their work is restricted to handling only "upward privacy breaches" (Evfimievski et al. 2003), whereas FRAPP handles downward privacy breaches as well.

Another model of privacy-preserving data mining is the k-anonymity model (Samarati and Sweeney 1998; Aggarwal and Yu 2004), where each record value is replaced with a corresponding generalized value. Specifically, each perturbed record cannot be distinguished from at least k other records in the data. However, the constraints of this model are less strict than ours since the intermediate database-forming-server can learn or recover precise records.

A different perspective is taken in Hippocratic databases, which are database systems that take responsibility for the privacy of the data they manage, and are discussed by Agrawal et al. (2002, 2004a,b), LeFevre et al. (2004). They involve specification of how the data is to be used in a privacy policy, and enforcing limited disclosure rules for regulatory concerns prompted by legislation.

Finally, the problem addressed by Atallah et al. (1999), Dasseni et al. (2001), Saygin et al. (2001, 2002) is preventing *sensitive models* from being inferred by the data miner—this work is complementary to ours since it addresses concerns about *output* privacy, whereas our focus is on the privacy of the *input* data. Maintaining input data privacy is considered by Kantarcioglu and Clifton (2002), Vaidya and Clifton (2002, 2003, 2004) in the context of databases that are *distributed* across a number of sites with each site only willing to share data mining results, but not the source data.

3 The FRAPP framework

In this section, we describe the construction of the FRAPP framework, and its quantification of privacy and accuracy measures.

Data model We assume that the original (true) database U consists of N records, with each record having M categorical attributes. The domain of attribute j is denoted by S_U^j, resulting in the domain S_U of a record in U being given by $S_U = \prod_{j=1}^{M} S_U^j$. We map the domain S_U to the index set $I_U = \{1, \ldots, |S_U|\}$, thereby modeling the database as a set of N values from I_U. If we denote the ith record of U as U_i, then $U = \{U_i\}_{i=1}^{N}, U_i \in I_U$.

To make this concrete, consider a database U with 3 categorical attributes *Age*, *Sex* and *Education* having the following category values:

Age	Child, Adult, Senior
Sex	Male, Female
Education	Elementary, Graduate

For this schema, $M = 3, S_U^1 = \{\text{Child, Adult, Senior}\}, S_U^2 = \{\text{Male, Female}\}, S_U^3 = \{\text{Elementary, Graduate}\}, S_U = S_U^1 \times S_U^2 \times S_U^3, |S_U| = 12$. The domain S_U is indexed by the index set $I_U = \{1, \ldots, 12\}$, and hence the set of records

U		
Child	Male	Elementary
Child	Male	Graduate
Child	Female	Graduate
Senior	Male	Elementary

maps to

U
1
2
4
9

Each record U_i represents the private information of customer i. Further, we assume that the U_i's are independent and identically distributed according to a fixed distribution p_U. This distribution p_U is not private and the customers are aware that the miner is expected to learn it—in fact, that is usually the *goal* of the data mining exercise. However, the assumption of independence implies that once p_U is known, possession of the private information U_j of any other customer j provides no additional inferences about customer i's private information U_i (Evfimievski et al. 2002).

Perturbation model As mentioned in Sect. 1, we consider the B2C privacy situation wherein the customers trust *no one except themselves*, that is, they wish to perturb their records at their client sites before the information is sent to the miner, or any intermediate party. This means that perturbation is carried out at the granularity of *individual* customer records U_i, without being influenced by the contents of the other records in the database.

For this situation, there are two possibilities: (a) A simple *independent attribute perturbation*, wherein the value of each attribute in the user record is perturbed independently of the rest; or (b) A more generalized *dependent attribute perturbation*, where the perturbation of each attribute may be affected by the perturbations of the other attributes in the record. Most of the prior perturbation techniques, including Evfimievski et al. (2002, 2003), Rizvi and Haritsa (2002), fall into the independent attribute perturbation category. The FRAPP framework, however, includes both kinds of perturbation in its analysis.

Let the perturbed database be $V = \{V_1, \ldots, V_N\}$, with domain S_V, and corresponding index set I_V. For example, given the sample database U discussed above, and assuming that each attribute is distorted to produce a value within its original domain, the distortion may result in

V
5
7
2
12

which maps to

V		
Adult	Male	Elementary
Adult	Female	Elementary
Child	Male	Graduate
Senior	Female	Graduate

Let the probability of an original customer record $U_i = u, u \in I_U$ being perturbed to a record $V_i = v, v \in I_V$ using randomization opertor $R(u)$ be $p(u \rightarrow v)$, and let A denote the matrix of these transition probabilities, with $A_{vu} = p(u \rightarrow v)$. This random process maps to a Markov process, and the perturbation matrix A should therefore satisfy the following properties (Strang 1988):

$$A_{vu} \geq 0 \quad \text{and} \quad \sum_{v \in I_V} A_{vu} = 1 \quad \forall u \in I_U,\ v \in I_V \tag{1}$$

Due to the constraints imposed by Eq. 1, the domain of A is a *subset* of $\mathbf{R}^{|S_V| \times |S_U|}$. This domain is further restricted by the choice of the randomization operator. For example, for the MASK technique (Rizvi and Haritsa 2002) mentioned in Sect. 1, all the entries of matrix A are decided by the choice of a single parameter, namely, the flipping probability.

In this paper, we explore the *preferred choices* of A to simultaneously achieve data privacy guarantees and high model accuracy, without restricting ourselves ab initio to a particular perturbation method.

3.1 Privacy guarantees

The miner is provided the perturbed database V, and the perturbation matrix A. Obviously, by receiving V_i corresponding to customer i, the miner gains partial information about U_i. However, as mentioned earlier in this section, due to the independence assumption, all V_i for $j \neq i$ disclose nothing about U_i—they certainly help the miner to learn the *distribution* p_U, but this is already factored in our privacy analysis since we assume the most conservative scenario wherein the miner has complete and precise knowledge of p_U. In fact, extracting information about p_U is typically the goal of the data mining exercise and, therefore, our privacy technique must encourage, rather than preclude, achieving this objective. The problem therefore reduces to analyzing specifically how much can be disclosed by V_i about the particular source record U_i.

We utilize the definition, given by Evfimievski et al. (2003), that a property $Q(u)$ of a data record $U(i) = u$ is a function $Q: u \rightarrow$ {true, false}. Further, a property holds for a record $U_i = u$ if $Q(u) =$ true. For example, consider the following record from our example dataset U

Age	Sex	Education
Child	Male	Elementary

Sample properties of this data record are

$$\begin{aligned} &Q_1(U_i) \equiv \text{“}Age = Child \textbf{ and } Sex = Male\text{”}, \\ \text{and} \quad &Q_2(U_i) \equiv \text{“}Age = Child \textbf{ or } \ Adult\text{”}. \end{aligned}$$

For this context, the *prior probability* of a property of a customer's private information is the likelihood of the property in the absence of *any* knowledge about the customer's private information. On the other hand, the *posterior probability* is the likelihood of the property given the perturbed information from the customer and the knowledge of the prior probabilities through reconstruction from the perturbed database. Specifically, the prior probability of any property $Q(U_i)$ is given by

$$\begin{aligned} P[Q(U_i)] &= \sum_{u:Q(u)} P[U_i = u] \\ &= \sum_{u:Q(u)} p_U(u) \end{aligned}$$

The posterior probability of any such property can be computed using Bayes formula

$$\begin{aligned} P[Q(U_i)|V_i = v] &= \sum_{u:Q(u)} P[U_i = u|V_i = v] \\ &= \sum_{u:Q(u)} \frac{P[U_i = u] \cdot p[u \rightarrow v]}{P[V_i = v]} \end{aligned}$$

As discussed by Evfimievski et al. (2003), in order to preserve the privacy of some property of a customer's private information, the posterior probability of that property should not be *unduly different* from the prior probability of the property for the customer. This notion of privacy is quantified by Evfimievski et al. (2003) through the following results, where ρ_1 and ρ_2 denote the prior and posterior probabilities, respectively:

Privacy breach An upward ρ_1-to-ρ_2 privacy breach exists with respect to property Q if $\exists v \in S_V$ such that

$$P[Q(U_i)] \leq \rho_1 \quad \text{and} \quad P[Q(U_i)|R(U_i) = v] \geq \rho_2.$$

Conversely, a downward ρ_2-to-ρ_1 privacy breach exists with respect to property Q if $\exists v \in S_V$ such that

$$P[Q(U_i)] \geq \rho_2 \quad \text{and} \quad P[Q(U_i)|R(U_i) = v] \leq \rho_1.$$

Amplification A randomization operator $R(u)$ is at most γ-amplifying for $v \in S_V$ if

$$\forall u_1, u_2 \in S_U \colon \frac{p[u_1 \rightarrow v]}{p[u_2 \rightarrow v]} \leq \gamma$$

where $\gamma \geq 1$ and $\exists u$: $p[u \rightarrow v] > 0$. Operator $R(u)$ is at most γ-amplifying if it is at most γ-amplifying for all qualifying $v \in S_V$.

Breach prevention Let R be a randomization operator, $v \in S_V$ be a randomized value such that $\exists u$: $p[u \rightarrow v] > 0$, and $\rho_1, \rho_2 (0 < \rho_1 < \rho_2 < 1)$ be two probabilities as per the above privacy breach definition. Then, if R is at most γ-amplifying for v, revealing "$R(u) = v$" will cause neither upward (ρ_1-to-ρ_2) nor downward (ρ_2-to-ρ_1) privacy breaches with respect to any property if the following condition is satisfied:

$$\frac{\rho_2(1 - \rho_1)}{\rho_1(1 - \rho_2)} > \gamma$$

If this situation holds, R is said to support (ρ_1, ρ_2) privacy guarantees.

From the above results of Evfimievski et al. (2003), we can derive for our formulation, the following condition on the perturbation matrix A in order to support (ρ_1, ρ_2) privacy:

$$\frac{A_{vu_1}}{A_{vu_2}} \leq \gamma < \frac{\rho_2(1 - \rho_1)}{\rho_1(1 - \rho_2)} \qquad \forall u_1, u_2 \in I_U, \forall v \in I_V \tag{2}$$

That is, the choice of perturbation matrix A should follow the restriction that the *ratio of any two matrix entries* (*in a row*) *should not be more than* γ.

Application environment At this juncture, we wish to clearly specify the environments under which the above guarantees are applicable. Firstly, our quantification of privacy breaches analyzes only the information leaked to the miner through observing the perturbed data; it does not take into account any prior knowledge that the miner may have about the original database. Secondly, we assume that the contents of each client's record are completely independent from those of other customers—that is, there are no *inter-transaction* dependencies. Due to this independence assumption, all the $R(U_j)$ for $j \neq i$ do not disclose anything about U_i and can therefore be ignored in privacy analysis; they certainly help the miner to learn the distribution of the original data, but in our analysis we have already assumed that this distribution is fully known by the miner. So the problem reduces to evaluating how much can be disclosed by $R(U_i)$ about U_i Evfimievski et al. (2003). We also hasten to add that we do not make any such restrictive assumptions about *intra-transaction* dependencies—in fact, the objective of association-rule mining is precisely to establish such dependencies.

3.2 Reconstruction model

We now move on to analyzing how the distribution of the original database is reconstructed from the perturbed database. As per the perturbation model, a client C_i with data record $U_i = u, u \in I_U$ generates record $V_i = v, v \in I_V$ with probability $p[u \rightarrow v]$. The generation event can be viewed as a Bernoulli trial with success probability $p[u \rightarrow v]$. If we denote the outcome of the ith Bernoulli trial by the random variable Y_v^i, the total number of successes Y_v in N trials is given by the sum of the N Bernoulli random variables:

$$Y_v = \sum_{i=1}^{N} Y_v^i \tag{3}$$

That is, the total number of records with value v in the perturbed database is given by Y_v.

Note that Y_v is the sum of N independent *but non-identical* Bernoulli trials. The trials are non-identical because the probability of success varies from trial i to trial j, depending on the values of U_i and U_j, respectively. The distribution of such a random variable Y_v is known as the Poisson-Binomial distribution (Wang 1993).

From Eq. 3, the expectation of Y_v is given by

$$E(Y_v) = \sum_{i=1}^{N} E(Y_v^i) = \sum_{i=1}^{N} P(Y_v^i = 1) \tag{4}$$

Using X_u to denote the number of records with value u in the original database, and noting that $P(Y_v^i = 1) = p[u \rightarrow v] = A_{vu}$ for $U_i = u$, we get

$$E(Y_v) = \sum_{u \in I_U} A_{vu} X_u \tag{5}$$

Let $X = [X_1 X_2 \ldots X_{|S_U|}]^T$, $Y = [Y_1 Y_2 \ldots Y_{|S_V|}]^T$. Then, the following expression is obtained from Eq. 5:

$$E(Y) = AX \tag{6}$$

At first glance, it may appear that X, the distribution of records in the original database (and the objective of the reconstruction exercise), can be directly obtained from the above equation. However, we run into the difficulty that the data miner does not possess $E(Y)$, but only *a specific instance* of Y, with which he has to approximate $E(Y)$.[3] Therefore, we resort to the following approximation to Eq. 6:

$$Y = A\widehat{X} \tag{7}$$

where X is estimated as $\widehat{X}$. This is a system of $|S_V|$ equations in $|S_U|$ unknowns, and for the system to be uniquely solvable, a necessary condition is that the space of the perturbed database is a superset of the original database (i.e. $|S_V| \geq |S_U|$). Further, if the inverse of matrix A exists, the solution of this system of equations is given by

$$\widehat{X} = A^{-1}Y \tag{8}$$

providing the desired estimate of the distribution of records in the original database. Note that this estimation is *unbiased* because $E(\widehat{X}) = A^{-1}E(Y) = X$.

3.3 Estimation error

To analyze the error in the above estimation process, we employ the following well-known theorem from linear algebra Strang (1988):

Theorem 1 *Given an equation of the form $Ax = b$ and that the measurement $\hat{b}$ of b is inexact, the relative error in the solution $\hat{x} = A^{-1}\hat{b}$ satisfies*

$$\frac{\parallel \hat{x} - x \parallel}{\parallel x \parallel} \leq c \frac{\parallel \hat{b} - b \parallel}{\parallel b \parallel}$$

where c is the condition number of matrix A.

For a positive-definite matrix, $c = \lambda_{max}/\lambda_{min}$, where λ_{max} and λ_{min} are the maximum and minimum eigen-values of matrix A, respectively. Informally, the condition number is a measure of the sensitivity of a matrix to numerical operations. Matrices with condition numbers near one are said to be *well-conditioned*, i.e. stable, whereas those with condition numbers much greater than one (e.g. 10^5 for a $5*5$ Hilbert matrix Strang 1988) are said to be *ill-conditioned*, i.e. highly sensitive.

From Eqs. 6 and 8, coupled with Theorem 1, we have

$$\frac{\parallel \widehat{X} - X \parallel}{\parallel X \parallel} \leq c \frac{\parallel Y - E(Y) \parallel}{\parallel E(Y) \parallel} \tag{9}$$

which means that the error in estimation arises from two sources: First, the sensitivity of the problem, indicated by the condition number of matrix A; and second, the deviation

[3] If multiple distorted versions of the database are provided, then $E(Y)$ is approximated by the observed average of these versions.

of Y from its mean, i.e. the deviation of perturbed database counts from their expected values, indicated by the variance of Y. In the following two sections, we mathematically determine how to reduce this error by: (a) appropriately choosing the perturbation matrix to minimize the condition number, and (b) identifying the minimum size of the database required to (probabilistically) bound the deviation within a desired threshold.

4 Perturbation matrix with minimum condition number

The perturbation techniques proposed in the literature primarily differ in their choices for perturbation matrix A. For example:

(1) *MASK*: The MASK (Rizvi and Haritsa 2002) randomization scheme uses a matrix A with

$$A_{vu} = p^k (1-p)^{M_b - k} \tag{10}$$

where M_b is the number of *boolean* attributes when each categorical attribute j is converted into $\mid S_U^j \mid$ boolean attributes, $(1-p)$ is the bit flipping probability for each boolean attribute, and k is the number of attributes with matching bits between the perturbed value v and the original value u.

(2) *Cut-and-paste*: The cut-and-paste (C&P) randomization operator (Evfimievski et al. 2002) employs a matrix A with

$$A_{vu} = \sum_{z=0}^{M} p_M[z] \cdot \sum_{q=max\{0,z+l_u-M,l_u+l_v-M_b\}}^{min\{z,l_u,l_v\}} \frac{\binom{l_u}{q}\binom{M-l_u}{z-q}}{\binom{M}{z}} \cdot \binom{M_b - l_u}{l_v - q} \rho^{(l_v-q)} (1-\rho)^{(M_b-l_u-l_v+q)} \tag{11}$$

where

$$p_M[z] = \sum_{w=0}^{min\{K,z\}} \binom{M-w}{z-w} \rho^{(z-w)} (1-\rho)^{(M-z)} \cdot \begin{cases} 1 - M/(K+1) & \text{if } w = M \;\&\; w < K \\ 1/(K+1) & \text{o.w.} \end{cases}$$

here l_u and l_v are the number of 1 bits in the original record u and its corresponding perturbed record v, respectively, while K and ρ are operator parameters.

To enforce strict privacy guarantees, the parameter settings for the above methods are bounded by the constraints, given in Eqs. 1 and 2, on the values of the elements of the perturbation matrix A. It turns out that for practical values of privacy requirements, the resulting matrix A for these previous schemes is extremely *ill-conditioned*—in fact, the condition numbers in our experiments were of the order of 10^5 and 10^7 for MASK and C&P, respectively.

Such ill-conditioned matrices make the reconstruction very sensitive to the variance in the distribution of the perturbed database. Thus, it is important to carefully choose the matrix A such that it is well-conditioned (i.e. has a low condition number). If we decide on a distortion method ab initio, as in the earlier techniques, then there is little room for making specific choices of perturbation matrix A. Therefore, we take the opposite approach of *first designing matrices of the required type*, and then devising perturbation methods that are compatible with these matrices.

To choose a suitable matrix, we start from the intuition that for $\gamma = \infty$, the obvious matrix choice is the *unity matrix*, which both satisfies the constraints on matrix A (Eqs. 1 and 2), and has the lowest possible condition number, namely, 1. Hence, for a given γ, we can choose the following matrix:

$$A_{ij} = \begin{cases} \gamma x & \text{if } i = j \\ x & \text{o.w.} \end{cases} \quad \text{where } x = \frac{1}{\gamma + (|S_U| - 1)} \tag{12}$$

which is of the form

$$x \begin{bmatrix} \gamma & 1 & 1 & \ldots \\ 1 & \gamma & 1 & \ldots \\ 1 & 1 & \gamma & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

It is easy to see that the above matrix, which incidentally is symmetric positive-definite and Toeplitz (Strang 1988), also satisfies the conditions given by Eqs. 1 and 2. Further, its condition number can be algebraically computed (as shown in the Appendix) to be $1 + \frac{|S_U|}{\gamma - 1}$. At an intuitive level, this matrix implies that the probability of a record u remaining as u after perturbation is γ times the probability of its being distorted to some $v \neq u$. For ease of exposition, we will hereafter informally refer to this matrix as the "Gamma-Diagonal matrix".

At this point, an obvious question is whether it is possible to design matrices with even lower condition number than the gamma-diagonal matrix. We prove next that the gamma-diagonal matrix has the *lowest* possible condition number among the class of symmetric positive-definite perturbation matrices satisfying the constraints of the problem, that is, it is an *optimal choice* (albeit non-unique).

4.1 Proof of optimality

Theorem 2 *Under the given privacy constraints, the Gamma-Diagonal matrix has the lowest condition number in the class of symmetric positive-definite perturbation matrices.*

Proof To prove this proposition, we will first derive the expression for minimum condition number of symmetric positive-definite matrices. For such matrices, the condition number is given by $c = \lambda_{max}/\lambda_{min}$, where λ_{max} and λ_{min} are the maximum and minimum eigen-values of the matrix, respectively. Further, since A is a Markov matrix (refer Eq. 1), the following results for eigen-values of a Markov matrix (Strang 1988) are applicable. □

Theorem 3 *For an $n \times n$ Markov matrix, one of the eigen-values is* 1*, and the remaining $n - 1$ eigen-values all satisfy* $\mid \lambda_i \mid \leq 1$.

Theorem 4 *The sum of the n eigen-values equals the sum of the n diagonal entries, that is,*

$$\lambda_1 + \cdots + \lambda_n = A_{11} + \cdots + A_{nn}$$

From Theorem 3, we obtain $\lambda_{max} = 1$, and from Theorem 4, that the sum of the rest of the eigen-values is fixed. If we denote $\lambda_1 = \lambda_{max}$, it is straightforward to see that λ_{min} is maximized when $\lambda_2 = \lambda_3 \cdots = \lambda_n$, leading to $\lambda_{min} = \frac{1}{n-1}\sum_{i=2}^{n} \lambda_i$. Therefore,

$$\lambda_{min} \leq \frac{1}{n-1}\sum_{i=2}^{n} \lambda_i$$

Using Theorem 4, we directly get

$$\lambda_{min} \leq \frac{1}{n-1}\left(\sum_{i=1}^{n} A_{ii} - 1\right)$$

resulting in the matrix condition number being lower-bounded by

$$c = \frac{1}{\lambda_{min}} \geq \frac{n-1}{\sum_{i=1}^{n} A_{ii} - 1} \tag{13}$$

Due to the privacy constraints on A given by Eq. 2,

$$A_{ii} \leq \gamma A_{ij} \qquad \forall j \neq i$$

Summing the above equation over all values of j except $j = i$, we get

$$\begin{aligned}(n-1)A_{ii} &\leq \gamma \sum_{j \neq i} A_{ij} \\ &= \gamma(1 - A_{ii})\end{aligned}$$

where the second step is due to the condition on A given by Eq. 1 and the restriction to symmetric positive-definite matrices. Solving for A_{ii} results in

$$A_{ii} \leq \frac{\gamma}{\gamma + n - 1} \tag{14}$$

and using this inequality in Eq. 13, we finally obtain

$$c \geq \frac{n-1}{\frac{n\gamma}{\gamma+n-1}-1} = \frac{\gamma+n-1}{\gamma-1} = 1 + \frac{n}{\gamma-1} \tag{15}$$

Therefore, the minimum condition number for the symmetric positive-definite perturbation matrices under privacy constraints represented by γ is $(1+\frac{n}{\gamma-1})$. The condition number of our "gamma-diagonal" matrix of size $|S_U|$ can be computed as shown in the Appendix, and its value turns out to be $(1+\frac{|S_U|}{\gamma-1})$. Thus, it is a *minimum condition number* perturbation matrix.

5 Database size and mining accuracy

In this section, we analyze the dependence of deviations of itemset counts in the perturbed database from their expected values, with respect to the size of the database. Then, we give bounds on the database sizes required for obtaining a desired accuracy.

As discussed earlier, Y_v denotes the total number of records with value v in the perturbed database, given by

$$Y_v = \sum_{1}^{N} Y_v^i$$

where Y_v^i is the Bernoulli random variable for record i, and N is the size of the database. To bound the deviation of Y_v from its expected value $E(Y_v)$, we use Hoeffding's General Bound (Motwani and Raghavan 1995), which bounds the deviation of the sum of Bernoulli random variables from its mean. Using these bounds for Y_v, we get

$$P\left(\frac{|\, Y_v - E(Y_v)\,|}{N} < \Delta\right) \geq 1 - 2e^{-2\Delta^2 N}$$

where Δ $(0 < \Delta < 1)$ represents the desired upper bound on the normalized deviation.

For the above probability to be greater than a user-specified value ϵ, the value of N should satisfy the following:

$$\begin{aligned} 1 - 2e^{-2\Delta^2 N} &\geq \epsilon \\ \Rightarrow N &\geq \ln(2/(1-\epsilon))/(2\Delta^2) \end{aligned} \tag{16}$$

That is, to achieve the desired accuracy (given by Δ), with the desired probability (given by ϵ), the miner must collect data from at least the number of customers given by the above bound. For example, with $\Delta = 0.001$ and $\epsilon = 0.95$, this turns out to be $N \geq 2 \times 10^6$, which is well within the norm for typical e-commerce environments. Moreover, note that these acceptable values were obtained with the Hoeffding Bound, a comparatively loose bound, and that in practice, it is possible that even datasets that do not fully meet this requirement may be capable of providing the desired accuracy.

For completeness, we now consider the hopefully rare situation wherein the customers are so few that accuracy cannot be guaranteed as per Eq. 16. Here, one approach that could be taken is to collect *multiple independent perturbations* of each customer's record, thereby achieving the desired target size. But, this has to be done carefully since the multiple distorted copies can potentially lead to a privacy breach, as described next.

5.1 Multiple versions of perturbed database

Assume that each user perturbs his/her record m times independently, so that overall the miner obtains m versions of the perturbed database. We hereafter refer to the set of perturbed records that share a common source record as "siblings".

Recall that a basic assumption made when defining privacy breaches in Sect. 3.1 was that the perturbed value $R(U_i)$ for the record i does not reveal *any* information about a record $j \neq i$. This assumption continues to be true in the multiple-versions variant if the miner is not aware of which records in the perturbed data set are siblings. Consequently, the privacy analysis of Sect. 3 can be applied verbatim to prove γ-amplification privacy guarantees in this environment as well. Therefore, all that needs to be done is to choose m such that the overall size of the database satisfies Eq. 16.

5.1.1 Multiple known siblings

The preceding analysis still leaves open the question as to what happens in situations wherein the data miner *is aware* of the siblings in the perturbed data set? It appears to us that maintaining accuracy requirements under such extreme circumstances may require relaxing the privacy constraints, as per the following discussion: With the gamma-diagonal matrix, the probability of a data value remaining unchanged is more than the probability of its being altered to any other value. Therefore, to guess the original value, the miner will obviously look for the value that appears the most number of times in the sibling records. For example, if 9 out of 10 versions of a given record have the identical perturbed value for an attribute, the miner knows with high probability the original value of that attribute. Clearly, in this case, one sibling reveals information about another sibling, violating the assumption required for γ-amplification privacy. At first glance, it might appear that this problem can be easily tackled by treating each group of siblings as a single multi-dimensional vector; but this strategy completely nullifies the original objective of having multiple versions to enhance accuracy. Therefore, in the remainder of this section, we quantitatively investigate the impact on privacy of having multiple *known* siblings in the database, with privacy now defined as the *probability of correctly guessing the original value.*

The first analysis technique that comes to mind is to carry out a hypothesis test—"the value seen the maximum number of times is indeed the true value"—using the χ^2 statistic. However, this test is not practical in our environment because of the extreme skewness of the distribution and the large cardinalities of the value domain. Therefore, we pursue the following alternate line of analysis: Consider a particular record with original (true) value u, which is independently perturbed m times, producing m perturbed record values. Let n_v be the number of times a perturbed value v appears in these m values, and let $\mathcal{R}$ be the random variable representing the value which is present the maximum number of times, i.e., $\mathcal{R} = i$ if $\forall i, n_i > n_j$. Then, the probability of correctly guessing $\mathcal{R} = u$ is

$$P(\mathcal{R} = u) = P(\wedge_{v \neq u}(n_u > n_v)) \quad \text{with} \sum n_i = m$$

Clearly, if u appears less than or equal to $L = \lfloor \frac{m}{|S_V|} \rfloor$ times, it cannot be the most frequent occurrence, since there must be another value v appearing at least $\lceil \frac{m}{|S_V|} \rceil$ times in the perturbed records. Hence, the probability of a correct guess satisfies the following inequality:

$$\begin{aligned} P(\mathcal{R} = u) &= 1 - P(M \neq u) \\ &\leq 1 - P(n_u \leq L) \\ &= 1 - \sum_{k=1}^{L} P(n_u = k) \\ &= 1 - \sum_{k=1}^{L} {}^{m}C_k \cdot p^k \cdot (1-p)^{m-k} \end{aligned} \tag{17}$$

where p is the probability $p[u \rightarrow u]$. The last step follows from the fact that n_u is a binomially distributed random variable.

Observe that $L = \lfloor \frac{m}{|S_V|} \rfloor \geq 0$, and hence the above inequality can be reduced to

$$\begin{aligned} P(\mathcal{R} = u) &\leq 1 - P(n_u = 0) \\ &= 1 - (1-p)^m \end{aligned}$$

For the gamma-diagonal matrix, $p = p[u \rightarrow u] = \gamma x$, resulting in the probability of a correct guess being

$$P(\mathcal{R} = u) \leq 1 - (1 - \gamma x)^m \tag{18}$$

The record domain size $|S_V|$ can be reasonably expected to be (much) greater than m in most database environments. This implies that the value of $p = \gamma x$ will usually be very small, leading to an acceptably low guessing probability.

A legitimate concern here is that the miner may try to guess the values of *individual* sensitive attributes (or a subset of such attributes) in a record, rather than its entire contents. To assess this possibility, let us assume that u and v, which were used earlier to denote values of complete records, now refer to a single attribute. As derived later in Sect. 8, for an attribute of domain size $|S_V^1|$, the probability $p[u \rightarrow u]$ is given by:

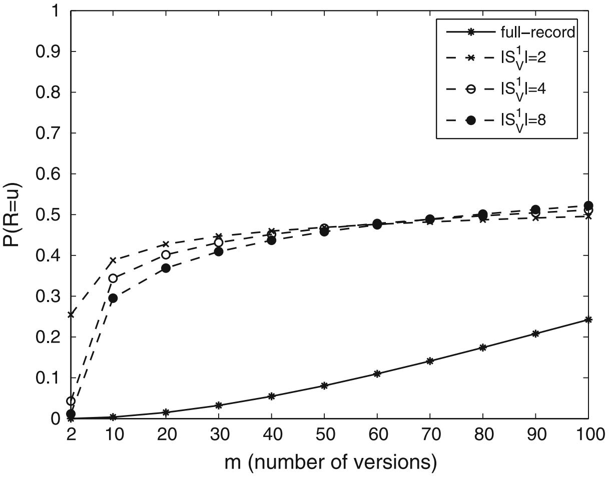

Fig. 1 $P(R = u)$ vs. m

$$p[u \rightarrow u] = \gamma x + \left(\frac{|S_V|}{|S_V^1|} - 1\right)x$$

An upper bound for the single-attribute guessing probability is directly obtained by substituting the above value of p, and $L = \lfloor \frac{m}{|S_V^1|} \rfloor$, in the inequality of Eq. 17.

A quantitative assessment of the number of versions that can be provided without jeopardizing user privacy is achieved by plotting the guessing probability upper bound against m, the number of versions. Sample plots are shown in Fig. 1 for a representative setup: $\gamma = 19$ with a record domain size $|S_V| = 2000$ and various single-attribute domain sizes ($S_V^1 = 2, 4, 8$). The solid line corresponds to the full-record whereas the dashed lines reflect the single-attribute cases.

Observe in Fig. 1 that the full-record guessing probability remains less than 0.1 even when the number of versions is as many as 50, and is limited to 0.25 for the extreme of 100 versions. Turning our attention to the single-attribute case, we see that for the lowest possible domain size, namely 2, the guessing probability levels off at around 0.5—note that this is no worse than the miner's ability to correctly guess the attribute value *without having access to the data*. Of course, for larger domain sizes such as 4 and 8, there is added information from the data—however, the key point again is that the guessing probabilities for these cases also level off around 0.5 in the practical range of m. In short, the miner's guess is at least *as likely to be wrong as it is to be correct*, which appears to be an acceptable privacy level in practice.

Moreover, for less stringent γ values, the guessing probabilities will decrease even further. Overall, these results imply that a substantial number of perturbed versions can be provided by users to the miner before their (guessing-probability) privacy can be successfully breached.

The observations in this section also indicate that FRAPP is robust against a potential privacy breach scenario where the information obtained from the users is (a) gathered periodically, (b) the set of users is largely the same, and (c) the data inputs of the users are often the same or very similar to their previous values. Such a scenario can occur, for example, when there is a core user community that regularly updates its subscription to an Internet service, like those found in the health or insurance industries. We therefore opine that FRAPP can be successfully used even in these challenging situations.

6 Randomizing the perturbation matrix

The estimation models discussed thus far implicitly assumed the perturbation matrix A to be *deterministic*. However, it appears intuitive that if the perturbation matrix parameters were themselves *randomized*, so that each client uses a perturbation matrix not specifically known to the miner, the privacy of the client will be further increased. Of course, it may also happen that the reconstruction accuracy suffers in this process. We explore this tradeoff, in this section, by replacing the deterministic matrix A with a randomized matrix $\tilde{A}$, where each entry $\tilde{A}_{vu}$ is a random variable with $E(\tilde{A}_{vu}) = A_{vu}$. The values taken by the random variables for a client C_i provide the specific parameter settings for her perturbation matrix.

6.1 Privacy guarantees

Let $Q(U_i)$ be a “property” (as explained in Sect. 3.1) of client C_i’s private information, and let record $U_i = u$ be perturbed to $V_i = v$. Denote the prior probability of $Q(U_i)$ by $P(Q(U_i))$. Then, on seeing the perturbed data, the posterior probability of the property is calculated to be:

$$\begin{aligned} P(Q(U_i)|V_i = v) &= \sum_{u:\,Q(u)} P_{U_i|V_i}(u|v) \\ &= \sum_{u:\,Q(u)} \frac{P_{U_i}(u) P_{V_i|U_i}(v|u)}{P_{V_i}(v)} \end{aligned}$$

When a deterministic perturbation matrix A is used for all clients, then $\forall i \quad P_{V_i|U_i}(v|u) = A_{vu}$, and hence

$$P(Q(U_i)|V_i = v) = \frac{\sum_{Q(u)} P_{U_i}(u) A_{vu}}{\sum_{Q(u)} P_{U_i}(u) A_{vu} + \sum_{\neg Q(u)} P_{U_i}(u) A_{vu}}$$

As discussed by Evfimievski et al. (2003), the data distribution P_{U_i} can, in the worst-case, be such that $P(U_i = u) > 0$ only if $\{u \in I_U | Q(u);\ A_{vu} = \max_{Q(u')} A_{vu'}\}$ or $\{u \in I_U | \neg Q(u);\ A_{vu} = \min_{\neg Q(u')} A_{vu'}\}$. For the deterministic gamma-diagonal matrix, $\max_{Q(u')} A_{vu'} = \gamma x$ and $\min_{\neg Q(u')} A_{vu'} = x$, resulting in

$$P(Q(U_i)|V_i = v) = \frac{P(Q(u)) \cdot \gamma x}{P(Q(u)) \cdot \gamma x + P(\neg Q(u)) x}$$

Since the distribution P_U is known through reconstruction, the above posterior probability can be determined by the miner. For example, if $P(Q(u)) = 5\%$, and $\gamma = 19$, the posterior probability works out to 50% for perturbation with the gamma-diagonal matrix.

But, in the randomized matrix case, where $P_{V_i|U_i}(v|u)$ is a realization of random variable $\tilde{A}$, only its distribution (and not the exact value for a given i) is known to the miner. This means that posterior probability computations like the one shown above cannot be made by the miner for a given record U_i. To make this concrete, consider a randomized matrix $\tilde{A}$ such that

$$\tilde{A}_{uv} = \begin{cases} \gamma x + r & \text{if } u = v \\ x - \frac{r}{|S_U|-1} & \text{o.w.} \end{cases} \tag{19}$$

where $x = \frac{1}{\gamma+|S_U|-1}$ and r is a random variable uniformly distributed between $[-\alpha, \alpha]$. Here, the worst-case posterior probability (and, hence, the privacy guarantee) for a record U_i is a function of the value of r, and is given by

$$\begin{aligned}\rho_2(r) &= P(Q(u)|v) \\ &= \frac{P(Q(u)) \cdot (\gamma x + r)}{P(Q(u)) \cdot (\gamma x + r) + P(\neg Q(u))(x - \frac{r}{|S_U|-1})}\end{aligned}$$

Therefore, only the posterior probability *range*, that is, $[\rho_2^-, \rho_2^+] = [\rho_2(-\alpha), \rho_2(+\alpha)]$, and the distribution over this range, can be determined by the miner. For example, for the scenario where $P(Q(u)) = 5\%$, $\gamma = 19$, and $\alpha = \gamma x/2$, the posterior probability lies in the range [33%, 60%], with its probability of being greater than 50% (ρ_2 at $r = 0$) equal to its probability of being less than 50%.

6.2 Reconstruction model

With minor modifications, the reconstruction model analysis for the randomized perturbation matrix $\tilde{A}$ can be carried out similar to that carried out earlier in Sect. 3.2 for the deterministic matrix A. Specifically, the probability of success for Bernoulli variable Y_v^i is now modified to

$$P(Y_v^i = 1|\tilde{A}_{vu}) = \tilde{A}_{vu}, \quad \text{for } U_i = u$$

and, from Eq. 4,

$$\begin{aligned}E(Y_v|\tilde{A}_{vu}) &= \sum_{i=1}^{N} P(Y_v^i = 1/\tilde{A}_{vu}) \\ &= \sum_{u \in I_U} \sum_{\{i|U_i=u\}} \tilde{A}_{vu}\end{aligned}$$

$$= \sum_{u \in I_U} \tilde{A}_{vu} X_u$$

$$\Rightarrow E(Y|\tilde{A}) = \tilde{A}X \tag{20}$$

leading to

$$E(E(Y|\tilde{A})) = AX \tag{21}$$

We estimate X as $\widehat{X}$ given by the solution of the following equation

$$Y = A\widehat{X} \tag{22}$$

which is an approximation to Eq. 21. From Theorem 1, the error in estimation is bounded by:

$$\frac{\parallel \widehat{X} - X \parallel}{\parallel X \parallel} \leq c \frac{\parallel Y - E(E(Y|\tilde{A})) \parallel}{\parallel E(E(Y|\tilde{A})) \parallel} \tag{23}$$

where c is the condition number of perturbation matrix $A = E(\tilde{A})$.

We now compare these bounds with the corresponding bounds of the deterministic case. Firstly, note that, due to the use of the randomized matrix, there is a *double expectation* for Y on the RHS of the inequality, as opposed to the single expectation in the deterministic case. Secondly, only the numerator is different between the two cases since we can easily show that $E(E(Y|\tilde{A})) = AX$. The numerator can be bounded by

$$\begin{aligned} &\parallel Y - E(E(Y|\tilde{A})) \parallel \\ &\quad = \parallel (Y - E(Y|\tilde{A})) + (E(Y|\tilde{A}) - E(E(Y|\tilde{A}))) \parallel \\ &\quad \leq \parallel Y - E(Y|\tilde{A}) \parallel + \parallel E(Y|\tilde{A}) - E(E(Y|\tilde{A})) \parallel \end{aligned}$$

Here, $\parallel Y - E(Y|\tilde{A}) \parallel$ is taken to represent the empirical variance of random variable Y_v. Since Y_v is, as discussed before, Poisson-Binomial distributed, its variance is given by (Wang 1993)

$$Var(Y_v|\tilde{A}) = N\overline{p}_v - \sum_i (p_v^i)^2 \tag{24}$$

where $\overline{p}_v = \frac{1}{N}\sum_i p_v^i$ and $p_v^i = P(Y_v^i = 1|\tilde{A})$.

It is easily seen (by elementary calculus or induction) that among all combinations $\{p_v^i\}$ such that $\sum_i p_v^i = n\overline{p}_v$, the sum $\sum_i (p_v^i)^2$ assumes its minimum value when all p_v^i are equal. It follows that, if the average probability of success $\overline{p}_v$ is kept constant, $Var(Y_v)$ assumes its maximum value when $p_v^1 = \cdots = p_v^N$. In other words, the variability of p_v^i, or *its lack of uniformity, decreases the magnitude of chance fluctuations* (Feller 1988). By using random matrix $\tilde{A}$ instead of deterministic A, we increase the variability of p_v^i (now p_v^i assumes variable values for all i), hence decreasing the fluctuation of Y_v from its expectation, as measured by its variance. In short,

$\| \; Y - E(Y|\tilde{A}) \; \|$ is likely to be decreased as compared to the deterministic case, thereby reducing the error bound.

On the other hand, the value of the second term: $\| \; E(Y|\tilde{A}) - E(E(Y|\tilde{A})) \; \|$, which depends upon the variance of the random variables in $\tilde{A}$, is now positive whereas it was 0 in the deterministic case. Thus, the error bound is increased by this term.

Overall, we have a *trade-off* situation here, and as shown later in our experiments of Sect. 9, the trade-off turns out such that the two opposing terms almost cancel each other out, making the error only *marginally worse than the deterministic case*.

7 Implementation of perturbation algorithm

Having discussed the privacy and accuracy issues of the FRAPP approach, we now turn our attention to the *implementation* of the perturbation algorithm described in Sect. 3. For this, we effectively need to generate for each $U_i = u$, a discrete distribution with PMF $P(v) = A_{vu}$ and CDF $F(v) = \sum_{i \le v} A_{iu}$, defined over $v = 1, \ldots, | \; S_V \; |$.

A straightforward algorithm for generating the perturbed record v from the original record u is the following

```
(1)  Generate $r \sim \mathcal{U}(0, 1)$
(2)  Repeat for $v = 1, \ldots, | S_V |$
         if $F(v-1) < r \le F(v)$
         return $V_i = v$
```

where $\mathcal{U}(0, 1)$ denotes uniform continuous distribution over [0, 1].

This algorithm, whose complexity is proportional to the *product* of the cardinalities of the attribute domains, will require $| \; S_V \; | /2$ iterations on average which can turn out to be very large. For example, with 31 attributes, each with two categories, this amounts to 2^{30} iterations per customer! We therefore present below an alternative algorithm whose complexity is proportional to the *sum* of the cardinalities of the attribute domains.

Specifically, to perturb record $U_i = u$, we can write

$$\begin{aligned} P(V_i; U_i = u) &= P(V_{i1}, \ldots, V_{iM}; u) \\ &= P(V_{i1}; u) \cdot P(V_{i2}|V_{i1}; u) \ldots P(V_{iM}|V_{i1}, \ldots, V_{i(M-1)}; u) \end{aligned}$$

where V_{ij} denotes the jth attribute of record V_i. For the perturbation matrix A, this works out to

$$\begin{aligned} P(V_{i1} = a; u) &= \sum_{\{v|v(1)=a\}} A_{vu} \\ P(V_{i2} = b|V_{i1} = a; u) &= \frac{P(V_{i2} = b, V_{i1} = a; u)}{P(V_{i1} = a; u)} \\ &= \frac{\sum_{\{v|v(1)=a \text{ and } v(2)=b\}} A_{vu}}{P(V_{i1} = a; u)} \\ &\quad \ldots \text{and so on} \end{aligned}$$

where $v(i)$ denotes the value of the ith attribute for the record with value v.

When A is chosen to be the gamma-diagonal matrix, and n_j is used to represent $\prod_{k=1}^{j} \mid S_U^k \mid$, we get the following expressions for the above probabilities after some simple algebraic manipulations:

$$\begin{aligned} P(V_{i1} = b; U_{i1} = b) &= \left(\gamma + \frac{n_M}{n_1} - 1\right) x \\ P(V_{i1} = b; U_{i1} \neq b) &= \frac{n_M}{n_1} x \end{aligned} \tag{25}$$

and for the jth attribute

$$\begin{aligned} &P(V_{ij} = b | V_{i1}, \ldots, V_{i(j-1)}; U_{ij} = b) \\ &= \begin{cases} \frac{(\gamma + \frac{n_M}{n_j} - 1)x}{\prod_{k=1}^{j-1} p_k} & \text{if } \forall k < j,\, V_{ik} = U_{ik} \\ \frac{(\frac{n_M}{n_j})x}{\prod_{k=1}^{j-1} p_k} & \text{o.w.} \end{cases} \end{aligned} \tag{26}$$

$$P(V_{ij} = b | V_{i1}, \ldots, V_{i(j-1)}; U_{ij} \neq b) = \frac{(\frac{n_M}{n_j})x}{\prod_{k=1}^{j-1} p_k}$$

where p_k is the probability that V_{ik} takes value a, given that a is the outcome of the random process performed for the kth attribute, i.e. $p_k = P(V_{ik} = a | V_{i1}, \ldots, V_{i(k-1)}; U_i)$.

The above perturbation algorithm takes M steps, one for each attribute. For the first attribute, the probability distribution of the perturbed value depends only on the original value for the attribute and is given by Eq. 25. For any subsequent column j, to achieve the desired random perturbation, we use as input both its original value and the *perturbed values* of the previous $j-1$ columns, and then generate the perturbed value for j as per the discrete distribution given in Eq. 26. This is an example of *dependent column perturbation*, in contrast to the independent column perturbations used in most of the prior techniques.

Note that even though the perturbation of a column depends on the perturbed values of previous columns, the columns can be perturbed in *any order*. Specifically, the probability distribution for each column perturbation, as given by Eqs. 25 and 26, gets modified accordingly so that the overall distribution for record perturbation remains the same.

Finally, to assess the complexity of the algorithm, it is easy to see that the maximum number of iterations for generating the jth discrete distribution is $|S_U^j|$, and hence the maximum number of iterations for generating a perturbed record is $\sum_j |S_U^j|$.

Remark The scheme presented above gives a general approach to ensure that the complexity is proportional to the sum of attribute cardinalities, for any choice of perturbation matrix. However, specifically for the gamma-diagonal matrix, a simpler algorithm could be used. Namely, with probability $x(\gamma - 1)$ return the original tuple, otherwise choose the value of each attribute in the perturbed tuple uniformly and independently.[4] In this special case, the algorithm is a generalization of Warner's classical randomized response technique (Warner 1965).

8 Application to mining tasks

To illustrate the utility of the FRAPP framework, we demonstrate in this section how it can be integrated in two representative mining tasks, namely *association rule mining*, which identifies interesting correlations between database attributes Agrawal and Srikant (1994), and *classification rule mining*, which produces class labeling rules for data records based on an initial training set (Mitchell 1997).

8.1 Association rule mining

The core computation in association rule mining is to identify "frequent itemsets", that is, itemsets whose support (i.e. frequency) in the database is in excess of a user-specified threshold sup_{min}. Eq. 8 can be *directly used* to estimate the support of itemsets containing all M categorical attributes. However, in order to incorporate the reconstruction procedure into bottom-up association rule mining algorithms such as Apriori (Agrawal and Srikant 1994), we need to also be able to estimate the supports of itemsets consisting of only a *subset* of attributes—this procedure is described next.

Let C denote the set of all attributes in the database, and C_s be a subset of these attributes. Each of the attributes $j \in C_s$ can assume one of the $|S_U^j|$ values. Thus, the number of itemsets over attributes in C_s is given by $I_{C_s} = \prod_{j \in C_s} |S_U^j|$. Let $\mathcal{L}, \mathcal{H}$ denote generic itemsets over this subset of attributes.

A user record *supports* the itemset $\mathcal{L}$ if the attributes in C_s take the values represented by $\mathcal{L}$. Let the support of $\mathcal{L}$ in the original and distorted databases be denoted by $sup_{\mathcal{L}}^U$ and $sup_{\mathcal{L}}^V$, respectively. Then,

$$sup_{\mathcal{L}}^V = \frac{1}{N} \sum_{v \text{ supports } \mathcal{L}} Y_v$$

where Y_v denotes the number of records in V with value v (refer Sect. 3.2). From Eq. 7, we know

$$Y_v = \sum_{u \in I_U} A_{vu} \widehat{X}_u$$

and therefore, using the fact that A is symmetric,

[4] Note that the notion of independence is with regard to the *perturbation* process, not the data distributions of the attributes.

$$\begin{aligned} sup^V_{\mathcal{L}} &= \frac{1}{N} \sum_{v \text{ supports } \mathcal{L}} \sum_{u} A_{vu} \widehat{X}_u \\ &= \frac{1}{N} \sum_{u} \widehat{X}_u \sum_{v \text{ supports } \mathcal{L}} A_{vu} \end{aligned}$$

Grouping the records u by the itemsets $\mathcal{H}$ that they support:

$$sup^V_{\mathcal{L}} = \frac{1}{N} \sum_{\mathcal{H}} \sum_{u \text{ supports } \mathcal{H}} \widehat{X}_u \sum_{v \text{ supports } \mathcal{L}} A_{vu} \tag{27}$$

Analyzing the term $\sum_{v \text{ supports } \mathcal{L}} A_{vu}$ in the above equation, we see that it represents the sum of the entries of column u in A over rows v that support itemset $\mathcal{L}$. Now, consider the columns u that support a given itemset $\mathcal{H}$. Note that due to the structure of the gamma diagonal matrix A, if $\mathcal{H} = \mathcal{L}$, then one diagonal entry is part of this sum, otherwise the summation involves only non-diagonal terms. Therefore, for all u that support a given itemset $\mathcal{H}$:

$$\sum_{v \text{ supports } \mathcal{L}} A_{vu} = \left\{ \begin{array}{ll} \gamma x + (\frac{I_C}{I_{C_s}} - 1)x & \text{if } \mathcal{H} = \mathcal{L} \\ \frac{I_C}{I_{C_s}} x & \text{o.w.} \end{array} \right\} := \mathcal{A}_{\mathcal{H}\mathcal{L}} \tag{28}$$

i.e. the probability of an itemset remaining the same after perturbation is $\frac{\gamma + I_C/I_{C_s} - 1}{I_C/I_{C_s}}$ times the probability of it being distorted to any other itemset.

Substituting in Eq. 27:

$$\begin{aligned} sup^V_{\mathcal{L}} &= \frac{1}{N} \sum_{\mathcal{H}} \mathcal{A}_{\mathcal{H}\mathcal{L}} \sum_{u \text{ supports } \mathcal{H}} \widehat{X}_u \\ &= \sum_{\mathcal{H}} \mathcal{A}_{\mathcal{H}\mathcal{L}} \quad \widehat{sup^U}_{\mathcal{H}} \end{aligned}$$

Thus, we can estimate the supports of itemsets over any subset C_s of attributes using the matrix $\mathcal{A}$ which is of much smaller dimension ($I_{C_s} \times I_{C_s}$) for small itemsets as compared to the original full matrix A.

A legitimate concern here might be that the matrix inversion could become time-consuming as we proceed to larger itemsets making I_{C_s} large. Fortunately, the inverse for this matrix has a simple closed-form expression:

Theorem 5 *The inverse of $\mathcal{A}$ is a matrix of order $n = I_{C_s}$ of the form $\mathcal{B} = \{\mathcal{B}_{ij}\colon 1 \leq i \leq n,\ 1 \leq j \leq n\}$, where*

$$\mathcal{B}_{ij} = \begin{cases} \delta y & \text{if } i = j \\ y & \text{o.w.} \end{cases}$$

with $\delta = -(\gamma + n - 2)$ and $y = -\frac{I_{C_s}}{I_C} \cdot \frac{1}{(\gamma - 1)}$

Proof As both $\mathcal{A}$ and $\mathcal{B}$ are square matrices of the same order, $\mathcal{AB}$ and $\mathcal{BA}$ are valid products. Also it can be trivially seen (by actual multiplication) that $\mathcal{AB} = \mathcal{BA} = \mathcal{I}$, where I is the identity matrix of order I_{C_s}. □

The above closed-form inverse can be directly used in the reconstruction process, greatly reducing both space and time resources. Specifically, the reconstruction algorithm can now be very simply written as:

for each $\mathcal{L}$ from 1 **to** n **do**

$$sup_{\mathcal{L}}^{U} = sup_{\mathcal{L}}^{V}\delta y + (N - sup_{\mathcal{L}}^{V})y \tag{29}$$

where N is the database cardinality, $sup_{\mathcal{L}}^{V}$ and $sup_{\mathcal{L}}^{U}$ are the perturbed and reconstructed frequencies, respectively, and n is the size of the index set, which is I_{C_s} for a subset of attributes and I_C for full-length itemsets.

Thus we can efficiently reconstruct the counts of itemsets over any subset of attributes without needing to construct the counts of complete records, and our scheme can be implemented efficiently on bottom-up association rule mining algorithms such as Apriori (Agrawal and Srikant 1994). Further, it is trivially easy to incorporate FRAPP even in *incremental* association rule mining algorithms such as DELTA (Pudi and Haritsa 2000) which operate periodically on changing historical databases, and use the results of previous mining operations to minimize the amount of work carried out during each new mining operation.

8.2 Classification rule mining

We now turn our attention to the task of classification rule mining. The primary input required for this process is the distribution of attribute values for each class in the training data. This input can be produced through the "ByClass" privacy-preserving algorithm enunciated by Agrawal and Srikant (2000), which partitions the training data by class label, and then separately distorts and reconstructs the distributions for the records corresponding to each class. After this reconstruction, an off-the-shelf classifier can be used to produce the actual classification rules. However, a complication that may arise in the privacy-preserving environment is that of *negative reconstructed frequencies*, described next.

8.2.1 Negative reconstructed frequencies

During the reconstruction process, it is sometimes possible that using the expressions given in Eq. 29, *negative reconstructed frequencies* may arise—this is because, given a large index set, it is possible that several indices may have little or no representation at all (i.e. low $sup_{\mathcal{L}}^{V}$), even after perturbation of the dataset. While this occurs for association rule mining too, it is not a problem there because such itemsets are automatically *pruned* due to the minimum support criterion. In the case of classification, however, negative frequencies pose difficulties because (a) they lack meaningful interpretation, and (b) classification techniques based on calculating logarithms of the itemset frequencies, such as decision tree classifiers (Quinlan 1993), now become infeasible.

To address this problem, we first set all negative reconstructed frequencies to zero and then uniformly scale down the positive frequencies such that their sum remains equal to the original dataset size. The rationale is that records corresponding to negative frequencies are scarce in the original dataset (i.e. "outliers") and can therefore be ignored without significant loss of accuracy. Further the scaling down of the positive frequencies is consistent with rule generation since classification techniques are based on *relative* frequencies or distributions, rather than absolute frequencies.

9 Performance evaluation

We move on, in this section, to quantitatively assessing the utility of the FRAPP approach with respect to the privacy and accuracy levels that it can provide for association rule mining and classification rule mining.

9.1 Association rule mining

9.1.1 Datasets

Two datasets, *CENSUS* and *HEALTH*, are used in our experiments, which are both derived from real-world repositories. Since it has been established in several sociological studies (e.g. Cranor et al. 1999; Westin 1999) that users typically expect privacy on only a few of the database fields—usually sensitive attributes such as health, income, etc.—our datasets also project out a representative subset of the columns in the original databases. The complete details of the datasets are given below:

CENSUS This dataset contains census information for about 50,000 adult American citizens, and is available from the UCI repository http://www.ics.uci.edu/~mlearn/mlsummary.html. We used three categorical (`native-country, sex, race`) attributes and three continuous (`age, fnlwgt, hours-per-week`) attributes from the census database in our experiments, with the continuous attributes partitioned into discrete intervals to convert them into categorical attributes. The specific categories used for these six attributes are listed in Table 1.

Table 1 CENSUS dataset

Attribute	Categories
Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Female, Male
Native-country	United-States, Other
Age	$[15-35)$, $[35-55)$, $[55-75)$, ≥ 75
Fnlwgt	$[0-1e5]$, $[1e5-2e5)$, $[1e5-3e5)$, $[3e5-4e5)$, $\geq 4e5$
Hours-per-week	$[0-20)$, $[20-40)$, $[40-60)$, $[60-80)$, ≥ 80

Table 2 HEALTH dataset

Attribute	Categories
INCFAM20 (family income)	Less than \$20, 000; \$20,000 or more
HEALTH (health status)	Excellent; Very good; Good; Fair; Poor
SEX (sex)	Male; Female
PHONE (has telephone)	Yes, phone number given; Yes, no phone number given; No
AGE (age)	$[0-20)$, $[20-40)$, $[40-60)$, $[60-80)$, $\geq 80)$
BDDAY12 (bed days in past 12 months)	$[0-7)$, $[7-15)$, $[15-30)$, $[30-60)$, ≥ 60
DV12 (Doctor visits in past 12 months)	$[0-7)$, $[7-15)$, $[15-30)$, $[30-60)$, ≥ 60

Table 3 Frequent itemsets for $sup_{min} = 0.02$

	Itemset length						
	1	2	3	4	5	6	7
CENSUS	19	102	203	165	64	10	–
HEALTH	23	123	292	361	250	86	12

HEALTH This dataset captures health information for over 100,000 patients collected by the US government http://dataferrett.census.gov. We selected 4 categorical and 3 continuous attributes from the dataset for our experiments. These attributes and their categories are listed in Table 2.

The association rule mining accuracy of our schemes on these datasets was evaluated for a user-specified minimum support of $sup_{min} = 2\%$. Table 3 gives the number of frequent itemsets in the datasets for this support threshold, as a function of the itemset length.

9.1.2 Multiple versions

In Sect. 5, Eq. 16 gave the number of data records required to obtain relative inaccuracy of less than Δ with a probability greater than ϵ. For $\Delta = 0.001$, and $\epsilon = 0.95$, this turned out to be $N \geq 2 \times 10^6$. Note that we need to consider small values of Δ, since the error given by Δ will be further amplified by the condition number, as indicated by Eq. 9 for relative error in reconstructed counts.

Since the datasets available to us were much smaller than the desired N, we resorted to scaling each dataset by a factor of 50 to cross the size threshold, by providing multiple distortions of each user record. As per the discussion in Sect. 5.1, such scaling does not result in any additional privacy breach if the miner has no knowledge of the sibling identities. Further, even when the miner does possess this knowledge, 50 versions was shown to retain an acceptable privacy level under the modified (guessing-probability) privacy definition. A useful side-effect of the dataset scaling is that it also ensures that our results are applicable to large disk-resident databases.

9.1.3 Performance metrics

We measure the performance of the system with regard to the accuracy that can be provided for a given privacy requirement specified by the user.

Privacy The (ρ_1, ρ_2) strict privacy measure from Evfimievski et al. (2003) is used as the privacy metric. We experimented with a variety of privacy settings—for example, varying ρ_2 from 30% to 50% while keeping ρ_1 fixed at 5%, resulting in γ values ranging from 9 to 19. The value of ρ_1 is representative of the fact that users typically want to hide uncommon values which set them apart from the rest, while a ρ_2 value of 50% indicates that the user can still plausibly deny any value attributed to him or her since it is equivalent to a random coin-toss attribution.

Accuracy We evaluate two kinds of mining errors, *Support Error* and *Identity Error*, in our experiments. The Support Error (μ) metric reflects the average relative error (in percent) of the reconstructed support values for those itemsets that are correctly identified to be frequent. Denoting the number of frequent itemsets by $|F|$, the reconstructed support by $\widehat{sup}$ and the actual support by sup, the support error is computed over all frequent itemsets as

$$\mu = \frac{1}{\mid F \mid} \Sigma_{f \in F} \frac{\mid \widehat{sup}_f - sup_f \mid}{sup_f} * 100$$

The Identity Error (σ) metric, on the other hand, reflects the percentage error in identifying frequent itemsets and has two components: σ^+, indicating the percentage of false positives, and σ^- indicating the percentage of false negatives. Denoting the reconstructed set of frequent itemsets with R and the correct set of frequent itemsets with F, these metrics are computed as

$$\sigma^+ = \frac{\mid R - F \mid}{\mid F \mid} * 100 \qquad \sigma^- = \frac{\mid F - R \mid}{\mid F \mid} * 100$$

9.1.4 Perturbation mechanisms

We present the results for FRAPP and representative prior techniques. For all the perturbation mechanisms, the mining of the distorted database was done using the Apriori (Agrawal and Srikant 1994) algorithm, with an additional support reconstruction phase at the end of each pass to recover the original supports from the perturbed database supports computed during the pass (Agrawal et al. 2004; Rizvi and Haritsa 2002).

Specifically, the perturbation mechanisms evaluated in our study are the following:

DET-GD This scheme uses the deterministic gamma-diagonal perturbation matrix A (Sect. 4) for perturbation and reconstruction. The perturbation was implemented using the techniques described in Sect. 7, and the equations of Sect. 8.1 were employed to construct the perturbation matrix used in each pass of Apriori.

RAN-GD This scheme uses the randomized gamma-diagonal perturbation matrix $\tilde{A}$ (Sect. 6) for perturbation and reconstruction. Though, in principle, any distribution can be used for $\tilde{A}$, here we evaluate the performance of uniformly distributed $\tilde{A}$ (as given by Eq. 19) over the entire range of the α randomization parameter (0 to γx).

MASK This is the perturbation scheme proposed in Rizvi and Haritsa (2002), intended for boolean databases and characterized by a single parameter $1 - p$, which determines the probability of an attribute value being flipped. In our scenario, the categorical attributes are mapped to boolean attributes by making each value of the category an attribute. Thus, the M categorical attributes map to $M_b = \sum_j \mid S_U^j \mid$ boolean attributes.

The flipping probability $1-p$ was chosen as the lowest value which could satisfy the privacy constraints given by Eq. 2. The constraint $\forall v$: $\forall u_1, u_2 : \frac{A_{vu_1}}{A_{vu_2}} \leq \gamma$ is satisfied for MASK (Rizvi and Haritsa 2002), if $\frac{p^{M_b}}{(1-p)^{M_b}} \leq \gamma$. But, for each categorical attribute, one and only one of its associated boolean attributes takes value 1 in a particular record. Therefore, all the records contain exactly M number of 1^s. Hence the ratio of two entries in the matrix cannot be greater than $\frac{p^{2M}}{(1-p)^{2M}}$ and the following condition is sufficient for the privacy constraints to be satisfied:

$$\frac{p^{2M}}{(1-p)^{2M}} \leq \gamma$$

The above equation was used to determine the appropriate value of p. For $\gamma = 19$ (corresponding to $(\rho_1, \rho_2) = (5\%, 50\%)$), this value turned out to be 0.439 and 0.448 for the CENSUS and HEALTH datasets, respectively.

C&P This is the Cut-and-Paste perturbation scheme proposed by Evfimievski et al. (2002), with algorithmic parameters K and ξ. To choose K, we varied K from 0 to M, and for each K, ξ was chosen such that the matrix (Eq. 11) satisfies the privacy constraints (Eq. 2). The results reported here are for the (K, ξ) combination giving the best mining accuracy, which for $\gamma = 19$, turned out to be $K = 3$ and $\xi = 0.494$.

9.1.5 Experimental results

For the CENSUS dataset, the support (μ) and identity (σ^-, σ^+) errors of the four perturbation mechanisms (DET-GD, RAN-GD, MASK, C&P) for $\gamma = 19$ are shown in Fig. 2, as a function of the length of the frequent itemsets (the performance of RAN-GD is shown for randomization parameter $\alpha = \gamma x/2$). The corresponding graphs for the HEALTH dataset are shown in Fig. 3. Note that the support error (μ) graphs are plotted on a *log-scale*. The detailed results are presented here for a representative privacy requirement of $(\rho_1, \rho_2) = (5\%, 50\%)$, which was also used by Evfimievski et al. (2003), and results in $\gamma = 19$. Similar performance trends were observed for the other practical values of γ, with the results for $\gamma = 13.28$ and $\gamma = 9$ on CENSUS dataset shown in Figs. 4 and 5, respectively.

In these figures, we first note that DET-GD performs, on an absolute scale, extremely well, the error being of the order of 10% for the longer itemsets. Further, its

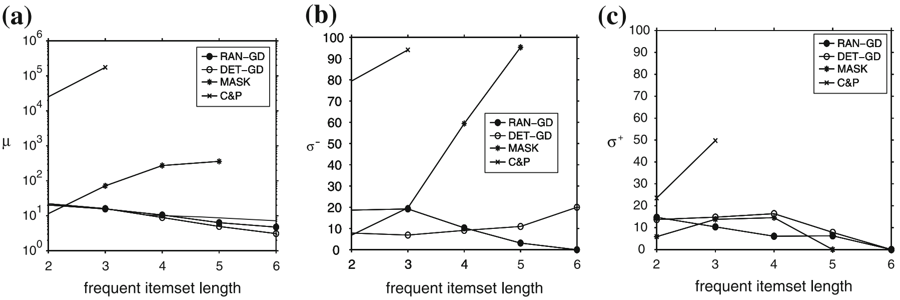

Fig. 2 CENSUS $\gamma = 19$. **a** Support error μ, **b** false negatives σ^-, **c** false positives σ^+

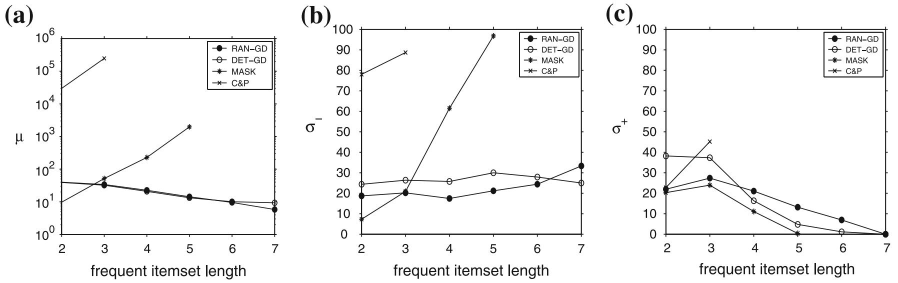

Fig. 3 HEALTH $\gamma = 19$. **a** Support error μ, **b** false negatives σ^-, **c** false positives σ^+

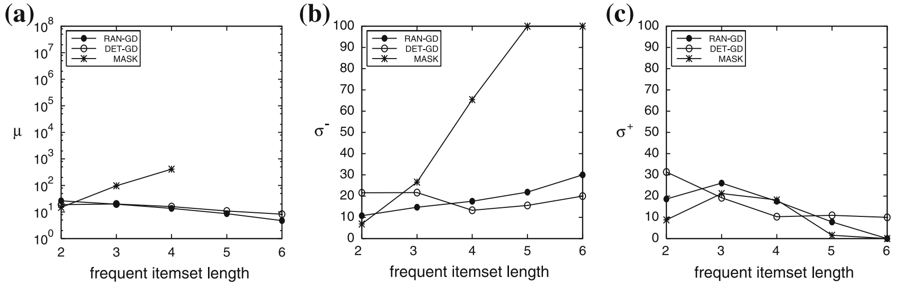

Fig. 4 Results for $\gamma = 13.28$ $(\rho_1, \rho_2) = (5\%, 41\%)$ on CENSUS. **a** Support error μ, **b** False negatives σ^-, **c** false positives σ^+

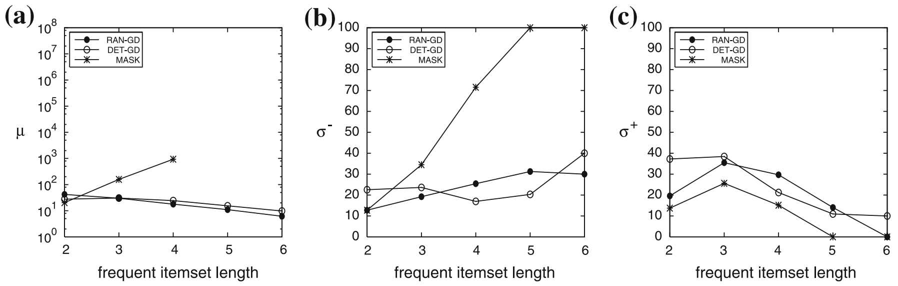

Fig. 5 Results for $\gamma = 9$ $(\rho_1, \rho_2) = (5\%, 32\%)$ on CENSUS. **a** Support error μ, **b** false negatives σ^-, **c** false positives σ^+

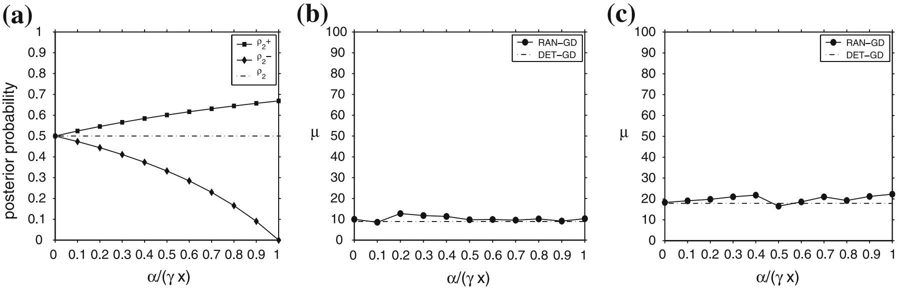

Fig. 6 Varying randomization of perturbation matrix ($\gamma = 19$). **a** Posterior probability, **b** support error μ (HEALTH), **c** support error μ (CENSUS)

performance is visibly better than that of MASK and C&P. In fact, as the length of the frequent itemset increases, the performance of both MASK and C&P degrade drastically. Specifically, MASK is not able to find any itemsets of length above 4 for the CENSUS dataset, and above 5 for the HEALTH dataset, while C&P could not identify itemsets beyond length 3 in both datasets.

The second point to note is that the accuracy of RAN-GD, although employing a randomized matrix, is only marginally worse than that of DET-GD. In return, it provides a substantial increase in the privacy—its worst-case (determinable) privacy breach is only 33% as compared to 50% with DET-GD. Figure 6a shows the performance of RAN-GD over the entire range of α with respect to the posterior probability range $[\rho_2^-, \rho_2^+]$. The mining support reconstruction errors for itemsets of length 4 are shown in Fig. 6b and c for the CENSUS and HEALTH datasets, respectively. We observe that the performance of RAN-GD does not deviate much from the deterministic case over the entire range, whereas very low *determinable* posterior probability is obtained for higher values of α.

Role of condition numbers The primary reason for DET-GD and RAN-GD's good performance is the low *condition numbers* of their perturbation matrices. This is quantitatively shown in Fig. 7, which plots these condition numbers on a *log-scale* (the condition numbers of DET-GD and RAN-GD are identical in this graph because $E(\tilde{A}) = A$). Note that the condition numbers are not only low but also *independent* of the frequent itemset length (algebraic computation of condition numbers is shown in the Appendix).

In marked contrast, the condition numbers for MASK and C&P increase *exponentially* with increasing itemset length, resulting in drastic degradation in accuracy. Thus, our choice of a gamma-diagonal matrix indicates highly promising results for discovery of long patterns.

Computational overheads Finally, with regard to actual mining response times also, FRAPP takes about *the same time* as Apriori for the complete mining process on the original and perturbed databases, respectively. This is because, as mentioned before, the reconstruction component shows up only *in between* mining passes and involves very simple computations (see Eq. 29). Further, the initial perturbation step took only a very modest amount of time even on vanilla PC hardware. Specifically, on a P-IV 2.0 GHz PC with 1 GB RAM and 40 GB hard disk, perturbing 2.5 million records of

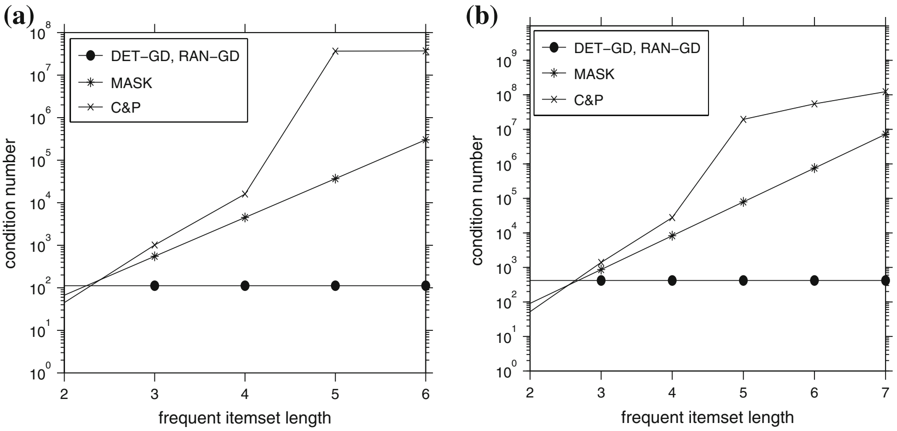

Fig. 7 Perturbation matrix condition numbers ($\gamma = 19$). **a** CENSUS, **b** HEALTH

CENSUS took about a minute, while 5 million records of HEALTH were distorted in a little over 2 min.

9.2 Classification rule mining

We now turn our attention to assessing the performance of FRAPP in the context of classification rule mining.

9.2.1 Experimental setup

The US Census dataset mentioned earlier was used in our experiments, out of which about 75% of the records were used for training and the remaining as test data. The attributes used in the experiment are given in Table 4, among which *salary* was chosen as the Class Label attribute. The classifier used is the highly popular public-domain C4.5 decision tree classifier Quinlan (1993), specifically the one available at http://www.cs.waikato.ac.nz/ml/weka.

Table 4 US CENSUS dataset for classification

Attribute	Categories
Native-country	United-States, Other
Salary	Less or equal to \$50, 000, Greater than \$50,000
Age	[15 − 35), [35 − 55), [55 − 75), ≥ 75
Type-of-employment	Private, Self Employment not Inc, Self Employment Inc, Federal Government, Local Government, State Government, Without pay, Never worked
Hours-per-week	[0 − 20), [20 − 40), [40 − 60), [60 − 80), ≥ 80

Table 5 Classification accuracy

Mining technique	Correct labeling (%)	Incorrect labeling (%)
FRAPP	72.88	27.12
DIRECT	75.34	24.66
BOTH	71.34	23.12

9.2.2 Experimental results

We choose a privacy level of $\gamma = 19$, corresponding to a maximum privacy breach of 50%. With this privacy setting, the accuracy results for FRAPP-based privacy-preserving classification are shown in Table 5, which also provides the corresponding accuracies for direct classification on the original database, representing in a sense, the "best case". We see here that the FRAPP accuracies are quite comparable to DIRECT, indicating that there is little cost associated with supporting the privacy functionality. Finally, the last line (BOTH) in Table 5 shows the proportion of cases where FRAPP and DIRECT *concurred* in their labeling—i.e. either both got it correct or both got it wrong, and as can be seen, the overlap between the two classifiers is very high, close to 95%.

10 Conclusions and future work

In this paper, we developed FRAPP, a generalized model for random-perturbation-based methods operating on categorical data under strict privacy constraints. The framework provides us with the ability to first make careful choices of the model parameters and then build perturbation methods for these choices. This results in order-of-magnitude improvements in model accuracy as compared to the conventional approach of deciding on a perturbation method upfront, which implicitly freezes the associated model parameters.

Using the framework, a "gamma-diagonal" perturbation matrix was identified as the best conditioned among the class of symmetric positive-definite matrices, and therefore expected to deliver the highest accuracy within this class. We also presented an implementation method for gamma-diagonal-based perturbation whose complexity is proportional to the sum of the domain cardinalities of the attributes in the database. Empirical evaluation of our approach on the CENSUS and HEALTH datasets demonstrated significant reductions in mining errors for association rule mining relative to prior privacy-preserving techniques, and comparable accuracy to direct mining for classification models.

The relationship between data size and model accuracy was also evaluated and it was shown that it is often possible to construct a sufficiently large dataset to achieve the desired accuracy by the simple expedient of generating multiple distorted versions of each customer's true data record, without materially compromising the data privacy.

Finally, we investigated the novel strategy of having the perturbation matrix composed of not values, but random variables instead. Our analysis of this approach indicated that at a marginal cost in accuracy, significant improvements in privacy levels could be achieved.

In our future work, we plan to investigate whether it is possible, as discussed in Sect. 2, to design distortion matrices such that the mining can be carried out directly on the distorted database without any explicit reconstruction—that is, to develop an "invariant FRAPP matrix".

Appendix

Condition number of gamma-diagonal matrix

We provide here the formula for computing the condition number of the gamma-diagonal distortion matrix. Specifically, consider the $n \times n$ matrix A of form

$$A_{ij} = \begin{cases} \xi x & \text{if } i = j \\ x & \text{o.w.} \end{cases}$$
$$\text{where } \xi x + (n-1)x = 1 \tag{30}$$

Since matrix A is symmetric, we can use the following well-known result (Strang 1988):

Theorem 6 *A symmetric matrix has real eigen-values.*

Let X be an eigenvector of the matrix A corresponding to eigenvalue λ. Then, it must satisfy:

$$AX = \lambda X$$

Using the structure of matrix A from Eq. 30, for any $i = 1, \ldots, n$,

$$\xi x X_i + \sum_{j \neq i} x X_j = \lambda X_i$$
$$\Leftrightarrow (\xi x - x) X_i + \sum_{j} x X_j = \lambda X_i$$

$$\Rightarrow \text{either } X_i = \frac{x \sum_j X_j}{\lambda + x - \xi x} \tag{31}$$

$$\text{or } \lambda + x - \xi x = 0 \tag{32}$$

Eq. 31 implies that all X_i are equal. Let this common value be g, leading to

$$\begin{aligned} g(\lambda + x - \xi x) &= ngx \\ \Rightarrow \lambda = \xi x + nx - x &= 1 \end{aligned} \tag{33}$$

From Eq. 32,

$$\lambda = (\xi - 1)x = \frac{(\xi - 1)}{(\xi + n - 1)} < 1$$

Thus, only two distinct values are taken by the eigen-values of matrix A: $\lambda_1 = 1$ and $\lambda_2 = \lambda_3 = \cdots = \lambda_n = \frac{(\xi-1)}{(\xi+n-1)}$. For $\xi \geq 1$, the eigen-values of the matrix A are positive, hence A is a positive-definite matrix, and its condition number is:

$$cond(A) = \frac{\lambda_{max}}{\lambda_{min}} = \frac{(\xi + n - 1)}{(\xi - 1)}$$

- For the matrix A given by Eq. 12, $\xi = \gamma, n =| S_U |, \gamma \geq 1$, so

$$cond(A) = \frac{(\gamma + \mid S_U \mid -1)}{(\gamma - 1)} = 1 + \frac{\mid S_U \mid)}{(\gamma - 1)}$$

- For matrix $\mathcal{A}$ for mining itemsets over subset of attributes C_s, given by Eq. 28, $\xi = \frac{\gamma + \frac{I_C}{I_{C_s}} - 1}{\frac{I_C}{I_{C_s}}}, n = I_{C_s}$.

  Hence,

$$\begin{aligned} \xi + n - 1 &= \frac{\gamma + I_C - 1}{\frac{I_C}{I_{C_s}}} \\ \xi - 1 &= \frac{\gamma - 1}{\frac{I_C}{I_{C_s}}} \\ cond(\mathcal{A}) &= \frac{(\xi + n - 1)}{(\xi - 1)} = \frac{(\gamma + I_C - 1)}{(\gamma - 1)} = 1 + \frac{\mid S_U \mid)}{(\gamma - 1)} \end{aligned}$$

References

Adam N, Wortman J (1989) Security control methods for statistical databases. ACM Comput Surv 21(4): 515–556

Aggarwal C, Yu P (2004, March) A condensation approach to privacy preserving data mining. In: Proceedings of the 9th international conference on extending database technology (EDBT), Heraklion, Crete, Greece

Agrawal D, Aggarwal C (2001, May) On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the ACM symposium on principles of database systems (PODS), Santa Barbara, California, USA

Agrawal R, Bayardo R, Faloutsos C, Kiernan J, Rantzau R, Srikant R (2004, August) Auditing compliance with a hippocratic database. In: Proceedings of the 30th international conference on very large data bases (VLDB), Toronto, Canada

Agrawal R, Kiernan J, Srikant R, Xu Y (2002, August) Hippocratic databases. In: Proceedings of the 28th international conference on very large data bases (VLDB), Hong Kong, China

Agrawal R, Kini A, LeFevre K, Wang A, Xu Y, Zhou D (2004, June) Managing healthcare data hippocratically. In: Proceedings of the ACM SIGMOD international conference on management of data, Paris, France

Agrawal R, Srikant R (1994, September) Fast algorithms for mining association rules. In: Proceedings of the 20th international conference on very large data bases (VLDB), Santiago de Chile, Chile

Agrawal R, Srikant R (2000, May) Privacy-preserving data mining. In: Proceedings of the ACM SIGMOD international conference on management of data, Dallas, Texas, USA

Agrawal R, Srikant R, Thomas D (2005, June) Privacy-preserving OLAP. In: Proceedings of the ACM SIGMOD international conference on management of data, Baltimore, Maryland, USA

Agrawal S, Krishnan V, Haritsa J (2004, March) On addressing efficiency concerns in privacy-preserving mining. In: Proceedings of the 9th international conference on database systems for advanced applications (DASFAA), Jeju Island, Korea

Atallah M, Bertino E, Elmagarmid A, Ibrahim M, Verykios V (1999, November) Disclosure limitation of sensitive rules. In: Proceedings of the IEEE knowledge and data engineering exchange workshop (KDEX), Chicago, Illinois, USA

Cranor L, Reagle J, Ackerman M (1999, April) Beyond concern: understanding net users' attitudes about online privacy, AT&T labs research technical report TR 99.4.3

Dasseni E, Verykios V, Elmagarmid A, Bertino E (2001, April) Hiding association rules by using confidence and support. In: Proceedings of the 4th international information hiding workshop (IHW), Pittsburgh, Pennsylvania, USA

de Wolf P, Gouweleeuw J, Kooiman P, Willenborg L (1998, March) Reflections on PRAM. In: Proceedings of the statistical data protection conference, Lisbon, Portugal

Denning D (1982) Cryptography and data security. Addison-Wesley

Duncan G, Pearson R (1991) Enhancing access to microdata while protecting confidentiality: prospects for the future. Stat Sci 6(3):219–232

Evfimievski A, Gehrke J, Srikant R (2003, June) Limiting privacy breaches in privacy preserving data mining. In: Proceedings of the ACM symposium on principles of database systems (PODS), San Diego, California, USA

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

Feller W (1988) An introduction to probability theory and its applications, vol I. Wiley

Gouweleeuw J, Kooiman P, Willenborg L, de Wolf P (1998) Post randomisation for statistical disclosure control: Theory and implementation. J Off Stat 14(4):485–502

Kantarcioglu M, Clifton C (2002, June) Privacy-preserving distributed mining of association rules on horizontally partitioned data. In: Proceedings of the ACM SIGMOD workshop on research issues in data mining and knowledge discovery (DMKD), Madison, Wisconsin, USA

Kargupta H, Datta S, Wang Q, Sivakumar K (2003, December) On the privacy preserving properties of random data perturbation techniques. In: Proceedings of the 3rd IEEE international conference on data mining (ICDM), Melbourne, Florida, USA

LeFevre K, Agrawal R, Ercegovac V, Ramakrishnan R, Xu Y, DeWitt D (2004, August) Limiting disclosure in hippocratic databases. In: Proceedings of the 30th international conference on very large data bases (VLDB), Toronto, Canada

Mishra N, Sandler M (2006, June) Privacy via pseudorandom sketches. In: Proceedings of the ACM symposium on principles of database systems (PODS), Chicago, Illinois, USA

Mitchell T (1997) Machine learning. McGraw Hill

Motwani R, Raghavan P (1995) Randomized algorithms. Cambridge University Press

Pudi V, Haritsa J (2000) Quantifying the utility of the past in mining large databases. Inf Sys 25(5):323–344

Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann

Rastogi V, Suciu D, Hong S (2007, September) The boundary between privacy and utility in data publishing. In: Proceedings of the 33rd international conference on very large data bases (VLDB), Vienna, Austria

Rizvi S, Haritsa J (2002, August) Maintaining data privacy in association rule mining. In: Proceedings of the 28th international conference on very large databases (VLDB), Hong Kong, China

Samarati P, Sweeney L (1998, June) Generalizing data to provide anonymity when disclosing information. In: Proceedings of the ACM symposium on principles of database systems (PODS), Seattle, Washington, USA

Saygin Y, Verykios V, Clifton C (2001) Using unknowns to prevent discovery of association rules. ACM SIGMOD Rec 30(4):45–54

Saygin Y, Verykios V, Elmagarmid A (2002, February) Privacy preserving association rule mining. In: Proceedings of the 12th international workshop on research issues in data engineering (RIDE), San Jose, California, USA

Shoshani A (1982, September) Statistical databases: characteristics, problems and some solutions. In: Proceedings of the 8th international conference on very large databases (VLDB), Mexico City, Mexico

Strang G (1988) Linear algebra and its applications. Thomson Learning Inc

Vaidya J, Clifton C (2002, July) Privacy preserving association rule mining in vertically partitioned data. In: Proceedings of the 8th ACM SIKGDD international conference on knowledge discovery and data mining (KDD), Edmonton, Alberta, Canada

Vaidya J, Clifton C (2003, August) Privacy-preserving k-means clustering over vertically partitioned data. In: Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining (KDD), Washington, DC, USA

Vaidya J, Clifton C (2004, April) Privacy preserving naive bayes classifier for vertically partitioned data. In: Proceedings of the SIAM international conference on data mining (SDM), Toronto, Canada

Wang Y (1993) On the number of successes in independent trials. Statistica Silica 3

Warner S (1965) Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc 60:63–69

Westin A (1999, July) Freebies and privacy: what net users think. Technical report, Opinion Research Corporation

Zhang N, Wang S, Zhao W (2004, September) A new scheme on privacy-preserving association rule mining. In: Proceedings of the 8th European conference on principles and practice of knowledge discovery in databases (PKDD), Pisa, Italy