A Framework for High-Accuracy Privacy-Preserving Mining

Shipra Agrawal Jayant R. Haritsa
Database Systems Lab, SERC/CSA
Indian Institute of Science, Bangalore 560012, INDIA
{shipra,haritsa}@dsl.serc.iisc.ernet.in

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 perturbation matrix with minimal condition number can be identified, maximizing the accuracy even under strict privacy guarantees. 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 cost in accuracy.

The quantitative utility of FRAPP, which applies to random-perturbation-based privacy-preserving mining in general, is evaluated specifically with regard to frequent-itemset mining on a variety of real datasets. Our experimental results indicate that, for a given privacy requirement, substantially lower errors are incurred, with respect to both itemset identity and itemset support, as compared to the prior techniques.

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.

To encourage users to submit correct inputs, a variety of privacy-preserving data mining techniques have been proposed in the last few years (e.g. [1, 5, 10, 14]). Their goal is to ensure the privacy of the raw local data but, at the same time, to 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 [14], intended for privacy-preserving association-rule mining on sparse boolean databases, each bit in the original 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 in [9], which guarantees strict limits on privacy breaches of individual user information, *independent of the distribution of the original data*. The distinguishing feature of FRAPP is its quantitative characterization of the *sources of error* in 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 privacy guarantees. In fact, we identify a perturbation matrix with provably minimal *condition number* (among the class of symmetric positive definite matrices), resulting in the best accuracy under the given constraints. An efficient implementation for this optimal perturbation matrix is also presented.

[1] Also the name of a popular coffee-based beverage, where the ingredients are perturbed and hidden under foam.

Proceedings of the 21st International Conference on Data Engineering (ICDE 2005)
1084-4627/05 $20.00 © 2005 **IEEE**

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 tradeoff is very attractive in that the privacy increase is substantial 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.

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 the popular mining task of identifying *frequent itemsets*, the cornerstone of association rule mining [3]. Our experiments on a variety of real datasets indicate that both identity and support errors are substantially lower than those incurred by the prior privacy-preserving techniques.

Another important difference with regard to the prior techniques is that their accuracy degrades with increasing itemset length, whereas FRAPP's accuracy is *robust* to this parameter. Therefore, it is particularly well-suited to datasets where the lengths of the maximal frequent itemsets are comparable to the attribute cardinality of the dataset.

1.2. Contributions

In a nutshell, the work presented here provides a mathematical foundation for "raising the bar, with respect to both accuracy and privacy, in strict privacy-preserving mining". 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 perturbation techniques for the proposed mechanisms;
- Quantitatively demonstrating the utility of our schemes in the context of association rule mining.

2. 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 database U consists of N independent and identically distributed 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 i^{th} 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={Child, Adult, Senior}, S_U^2={Male, Female}, S_U^3={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 $I_U = \{1, ..., 12\}$, and hence the set of tuples

U				U
Child	Male	Elementary		1
Child	Male	Graduate	maps	2
Child	Female	Graduate	to	4
Senior	Male	Elementary		9

Perturbation Model. We consider the 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 done 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 [9, 10, 14], 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 $V_i = v, v \in I_V$ 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 [15]:

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

Due to the constraints imposed by Equation 1, the domain of A is a *subset* of $\mathbf{R}^{|S_U| \times |S_V|}$. This domain is further restricted by the choice of perturbation method. For example, for the MASK technique [14] mentioned in the Introduction, 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 privacy guarantees and high accuracy, without restricting ourselves *ab initio* to a particular perturbation method.

2.1. Privacy Guarantees

The miner receives the perturbed database V, and the perturbation matrix A, and attempts to reconstruct the original probability *distribution* of database U. In 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. As discussed in [9], 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* to that of the prior probability of the property for the customer. This notion of privacy is quantified in [9] 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

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

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 suitable $v \in S_V$.

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 $0 < \rho_1 < \rho_2 < 1$ be two probabilities as per the privacy breach definition above. 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 [9], we can derive for our formulation, the following condition on the perturbation matrix A in order to support (ρ_1, ρ_2) privacy:

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

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

2.2. Reconstruction Model

We now analyze 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 $p[u \rightarrow v]$. This event of generation of v can be viewed as a Bernoulli trial with success probability $p[u \rightarrow v]$. If we denote the outcome of the i^{th} 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 [16].

From Equation 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 $X = [X_1 X_2 \cdots X_{|S_U|}]^T$, $Y = [Y_1 Y_2 \cdots Y_{|S_V|}]^T$. Then, the following expression is obtained from Equation 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 $E(Y)$.[2] Therefore, we resort to the following approximation to Equation 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. 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

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

[2] If multiple distorted versions happen to be provided, then $E(Y)$ is approximated by the observed average of these versions.

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

2.3. Estimation Error

To analyze the error in the above estimation process, the following well-known theorem from linear algebra [15] comes in handy:

Theorem 2.1 *Given an equation of the form $Ax = b$ and that the measurement of b is inexact, the relative error in the solution $x = A^{-1}b$ satisfies*

$$\frac{\| \delta x \|}{\| x \|} \leq c \frac{\| \delta b \|}{\| b \|}$$

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 eigenvalues 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 [15]) are said to be *ill-conditioned*, i.e. highly sensitive.

From Equations 6, 8 and Theorem 2.1, we have

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

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 of Y from its mean, indicated by the variance of Y.

As discussed previously, Y_v is a Poisson-Binomial distributed random variable. Using the well-known expression for variance of a Poisson-Binomial random variable [16], the variance of Y_v is computed to be

$$Var(Y_v) = A_v X (1 - \frac{1}{N} A_v X) - \sum_{u \in I_U} (A_{vu} - \frac{1}{N} A_v X)^2 X_u$$

which depends on the perturbation matrix A and the distribution X of records in the original database. Thus the effectiveness of the privacy preserving method is *critically dependent on the choice of matrix A.*

3. Choice of Perturbation Matrix

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

- MASK [14] 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.

- The *cut-and-paste* (C&P) randomization operator [10] employs a matrix A with

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

where

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

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 choice of listed parameters for the above methods are bounded by the constraints, given in Equations 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 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 apriori, 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 (Equations 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

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

It is easy to see that the above matrix, which incidentally is symmetric and Toeplitz [15], also satisfies the conditions given by Equations 1 and 2. Further, its condition number can be algebraically computed to be $1 + \frac{\mid S_U \mid}{\gamma - 1}$. 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 that have even lower condition number than the gamma-diagonal matrix. In [6], we prove that the gamma-diagonal matrix has the *lowest* possible condition number among the class of symmetric perturbation matrices satisfying the constraints of the problem, that is, it is an *optimal choice* (albeit non-unique).

4. Randomizing the Perturbation Matrix

The estimation model in the previous section 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 that is 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.

In this section, we explore this tradeoff, by replacing the deterministic matrix A with 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.

4.1. Privacy Guarantees

Let $Q(U_i)$ be a property 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:

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

When a deterministic perturbation matrix A is used for all clients, then $\forall i \;\; P_{V_i|U_i}(v|u) = A_{vu}$, 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 in [9], the data distribution P_{U_i} can, in the worst case, be such that $P(U_i = u) > 0$ only if $\{u \in I_U | Q(u)$ and $A_{vu} = \max_{Q(u')} A_{vu'}\}$ or $\{u \in I_U | \neg Q(u)$ and $A_{vu} = \min_{\neg Q(u')} A_{vu'}\}$. 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

$$\tilde{A}_{uv} = \begin{cases} \gamma x + r & \text{if } u = v \\ x - \frac{r}{|S_U|-1} & \text{o.w.} \end{cases} \tag{13}$$

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 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*, i.e. $[\rho_2^-, \rho_2^+] = [\rho_2(-\alpha), \rho_2(+\alpha)]$, 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%.

4.2. Reconstruction Model

With minor modifications, the reconstruction model analysis for the randomized perturbation matrix $\tilde{A}$ can be carried out similar to that done earlier in Section 2.2 for the deterministic matrix A (the complete details are available in [6]). At the end of this analysis, we find that the estimation error is bounded by

$$\frac{\| \widehat{X} - X \|}{\| X \|} \leq c \frac{\| Y - E(E(Y|\tilde{A})) \|}{\| E(E(Y|\tilde{A})) \|} \tag{14}$$

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

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

Here, $\| Y - E(Y|\tilde{A}) \|$ denotes the variance of random variable Y. Since Y_v, as discussed before, is Poisson-Binomial distributed, its variance is given by [16]

$$Var(Y_v|\tilde{A}) = N\overline{p}_v - \sum_i (p_v^i)^2 \tag{15}$$

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* [11]. 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 $\tilde{A}$, was 0 in the deterministic case, but is now positive. Thus, the error bound is increased by this term.

Overall, we have a *tradeoff* situation here, and as shown later in our experiments of Section 7, the tradeoff turns out such that the two opposing terms almost cancel each other out, making the error only *marginally worse than the deterministic case*.

5. Implementation of Perturbation Algorithm

To implement the perturbation process discussed in the previous sections, we effectively need to generate for each $U_i = u$, a discrete distribution with PMF $P(v) = A_{vu}$ and CDF $F(v) = \sum_{i \leq v} A_{iu}$, defined over $v = 1, \ldots, \mid S_V \mid$. While a direct implementation results in algorithmic complexity that is proportional to the *product* of the cardinalities of the attribute domains (see [6] for details), we present below an efficient 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
$P(V_i; U_i = u)$
$= P(V_{i1}, \ldots, V_{iM}; u)$
$= P(V_{i1}; u) \cdot P(V_{i2}|V_{i1}; u) \cdots P(V_{iM}|V_{i1}, \ldots, V_{i(M-1)}; u)$
where V_{ij} denotes the j^{th} attribute of record V_i. For the perturbation matrix A, this works out to

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

where $v(i)$ denotes the value of the i^{th} 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:

$$\begin{aligned} P(V_{i1} = b; U_{i1} = b) &= (\gamma + \frac{n_M}{n_1} - 1)x \\ P(V_{i1} = b; U_{i1} \neq b) &= \frac{n_M}{n_1} x \end{aligned} \quad (16)$$

and for the j^{th} attribute

$$P(V_{ij} = b|V_{i1}, \ldots, V_{i(j-1)}; U_{ij} = b) = \begin{cases} \frac{(\gamma + \frac{n_M}{n_j} - 1)x}{\prod_{k=1}^{j-1} p_k} & \text{if } \forall k < j, V_{ik} = U_{ik} \\ \frac{(\frac{n_M}{n_j})x}{\prod_{k=1}^{j-1} p_k} & \text{o.w.} \end{cases} \quad (17)$$

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

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 k^{th} attribute, i.e. $p_k = P(V_{ik} = a|V_{i1}, \ldots, V_{i(k-1)}; U_i)$.

Note that the above perturbation 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 Equation 16. 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 generate the perturbed value as per the discrete distribution given in Equation 17. This is an example of *dependent column perturbation*, in contrast to the independent column perturbations used in most of the prior techniques.

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

6. Application to Association Rule Mining

To illustrate the utility of the FRAPP framework, we demonstrate in this section how it can be used for enhancing privacy-preserving mining of *association rules*, a popular mining model that identifies interesting correlations between database attributes [4].

The core computation of association rule mining is to identify "frequent itemsets", that is, all those itemsets whose support (i.e. frequency) in the database is in excess of a user-specified threshold sup_{min}. Equation 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* [4], we need to also be able to estimate the supports of itemsets consisting of only a *subset* of attributes.

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 itemsets over this subset of attributes. Given this, we can show (details in [6])

that the matrix required to estimate supports of the subsets is

$$\mathcal{A}_{\mathcal{HL}} = \begin{cases} \gamma x + (\frac{I_C}{I_{C_s}} - 1)x & \text{if } \mathcal{H} = \mathcal{L} \\ \frac{I_C}{I_{C_s}} x & \text{o.w.} \end{cases} \quad (18)$$

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.

Using the above $I_{C_s} \times I_{C_s}$ matrix, the supports of itemsets over any subset C_s of attributes can be estimated. However, the inversion of the matrix could be potentially time-consuming if I_{C_s} is large. Fortunately, as described in [6], the following alternative equation can be derived and used to efficiently compute the support of a given itemset $\mathcal{H}$, since it only involves the inversion of a 2-by-2 matrix,

$$\mathcal{A}^{2\times 2} = x \begin{bmatrix} \gamma + (\frac{I_C}{I_{C_s}} - 1) & \frac{I_C}{I_{C_s}} \\ (I_{C_s} - 1)\frac{I_C}{I_{C_s}} & \gamma + (\frac{I_C}{I_{C_s}} - 1) + (I_{C_s} - 2)\frac{I_C}{I_{C_s}} \end{bmatrix}$$

with

$$\begin{bmatrix} sup^V_{\mathcal{H}} \\ N - sup^V_{\mathcal{H}} \end{bmatrix} = \mathcal{A}^{2\times 2} \begin{bmatrix} sup^U_{\mathcal{H}} \\ N - sup^U_{\mathcal{H}} \end{bmatrix}$$

where $sup^U_{\mathcal{H}}$ and $sup^V_{\mathcal{H}}$ denote the supports of itemset $\mathcal{H}$ in the original and perturbed databases, respectively. The above equation is derived using the fact that the sum of the supports of all the I_{C_s} itemsets is constrained to be equal to the total number of records in the database, i.e. N.

Hence, our scheme can be implemented efficiently on bottom-up association rule mining algorithms such as Apriori [4].

7. Performance Analysis

In this section, we quantitatively assess the utility of the FRAPP framework with respect to the privacy and accuracy levels that it can provide for mining frequent itemsets.

7.1. Datasets

The following real-world datasets were used in our experiments:

CENSUS. This dataset contains census information for approximately 50,000 adult American citizens, and is available from the UCI repository [18]. It includes fields that users may prefer to keep private – for example, the "race" and "sex" attributes. We used three continuous (`age, fnlwgt, hours-per-week`) and three categorical (`native-country, sex, race`) 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.

HEALTH. This dataset captures health information for over 100,000 patients collected by the US government [17]. We selected 3 continuous and 4 categorical attributes from the dataset for our experiments. The attributes and their categories are listed in Table 2.

In order to ensure that our results were applicable to large disk-resident databases, the above datasets were scaled by a factor of 50 in our experiments. We evaluated the association rule mining accuracy of our schemes on the above datasets 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.

7.2. Performance Metrics

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

Privacy. The (ρ_1, ρ_2) strict privacy measure from [9] is used as the privacy metric. While we experimented with a variety of privacy settings, due to space limitations, results are presented here for a sample $(\rho_1, \rho_2) = (5\%, 50\%)$ requirement, which was also used in [9]. This privacy setting results in $\gamma = 19$.

Accuracy. We evaluate two kinds of mining errors, **Support Error** and **Identity Error**, in our experiments. The Support Error (ρ) metric reflects the (percentage) average relative error in 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

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

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

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

Table 1. CENSUS Dataset

Attribute	*Categories*
age	$[15-35), [35-55), [55-75), \geq 75$
fnlwgt	$[0-1e5], [1e5-2e5), [1e5-3e5), [3e5-4e5), \geq 4e5$
hours-per-week	$[0-20), [20-40), [40-60), [60-80), \geq 80$
race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
sex	Female, Male
native-country	United-States, Other

Table 2. HEALTH Dataset

Attribute	*Categories*
AGE (Age)	$[0-20), [20-40), [40-60), [60-80), \geq 80)$
BDDAY12 (Bed days in past 12 months)	$[0-7), [7-15), [15-30), [30-60), \geq 60$
DV12 (Doctor visits in past 12 months)	$[0-7), [7-15), [15-30), [30-60), \geq 60$
PHONE (Has Telephone)	Yes, phone number given; Yes, no phone number given; No
SEX (Sex)	Male; Female
INCFAM20 (Family Income)	Less than $20,000; $20,000 or more
HEALTH (Health status)	Excellent; Very Good; Good; Fair; Poor

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

7.3. Perturbation Mechanisms

We present the results for FRAPP and representative prior techniques. For all the perturbation mechanisms, the mining from the distorted database was done using *Apriori* [4] 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 [7, 14].

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

DET-GD. This scheme uses the deterministic gamma-diagonal perturbation matrix A (Section 3) for perturbation and reconstruction. The implementation described in Section 5 was used to carry out the perturbation, and the equations of Section 6 were used to construct the perturbation matrix used in each pass of Apriori.

RAN-GD. This scheme uses the randomized gamma-diagonal perturbation matrix $\tilde{A}$ (Section 4) 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 Equation 13) over the entire range of the α randomization parameter $(0$ to $\gamma x)$.

MASK. This is the perturbation scheme proposed in [14], 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 Equation 2 (details of the procedure are given in [6]). For $\gamma = 19$, 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 in [10], with algorithmic parameters K and ρ. To choose K, we varied K from 0 to M, and for each K, ρ was chosen such that the matrix (Equation 11) satisfies the privacy constraints (Equation 2). The results reported here are for the (K, ρ) combination giving the best mining accuracy, which for $\gamma = 19$, turned out to be $K = 3$ and $\rho = 0.494$.

7.4. Experimental Results

For the CENSUS dataset, the support (ρ) and identity (σ^-, σ^+) errors of the four perturbation mechanisms (DET-

GD, RAN-GD, MASK, C&P) are shown in Figure 1, 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 Figure 2. Note that the support error (ρ) graphs are plotted on a *log-scale*.

In these figures, we first note that the performance of the DET-GD method 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 lower 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 3a 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 Figures 3b and 3c 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.

The primary reason for DET-GD and RAN-GD's good performance are the low *condition numbers* of their perturbation matrices. This is quantitatively shown in Figure 4, 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.

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.

8. Related Work

The issue of maintaining privacy in data mining has attracted considerable attention in the recent past. The work closest to our approach is that of [2, 5, 8, 9, 10, 13, 14]. In the pioneering work of [5], privacy-preserving data classifiers based on adding noise to the record values were proposed. This approach was extended in [2] and [13] to address a variety of subtle privacy loopholes.

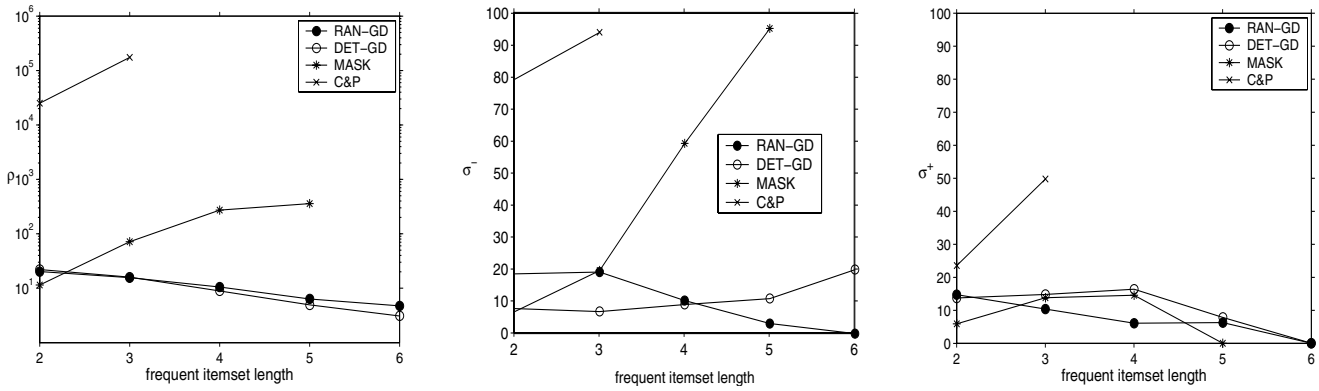
New randomization operators for maintaining data privacy for boolean data were presented and analyzed in [10, 14]. These methods are applicable to categorical/boolean data and are based on probabilistic mapping from the domain space to the range space, rather than by incorporating additive noise to continuous valued data. A theoretical formulation of privacy breaches for such methods, and a methodology for limiting them, were given in the foundational work of [9].

Techniques for probabilistic perturbation have also been investigated in the statistics literature. For example, the PRAM method [8, 12], 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 is certainly an attractive notion, the systematic identification of such matrices and the conditions on their applicability is still an open research issue.

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, in a sense, to 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.

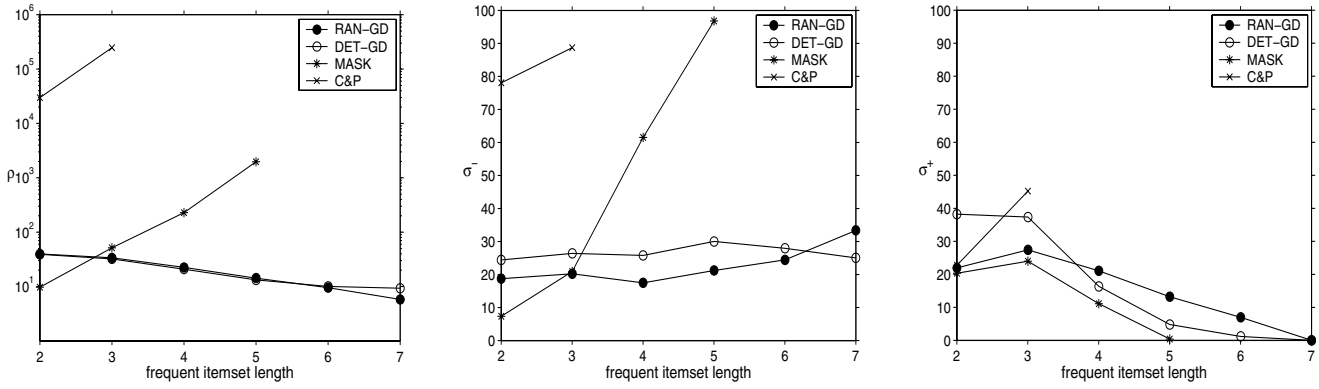
9. Conclusions and Future Work

In this paper, we developed FRAPP, a generalized model for random-perturbation-based methods operating on categorical data under strict privacy constraints. We showed that by making careful choices of the model parameters and building perturbation methods for these choices, order-of-magnitude improvements in accuracy could be achieved as compared to the conventional approach of first deciding on a method and thereby implicitly fixing the associated model parameters. In particular, a "gamma-diagonal" perturbation matrix was identified as delivering the best accuracy among the class of symmetric positive definite matrices. We presented an implementation technique 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 new gamma-



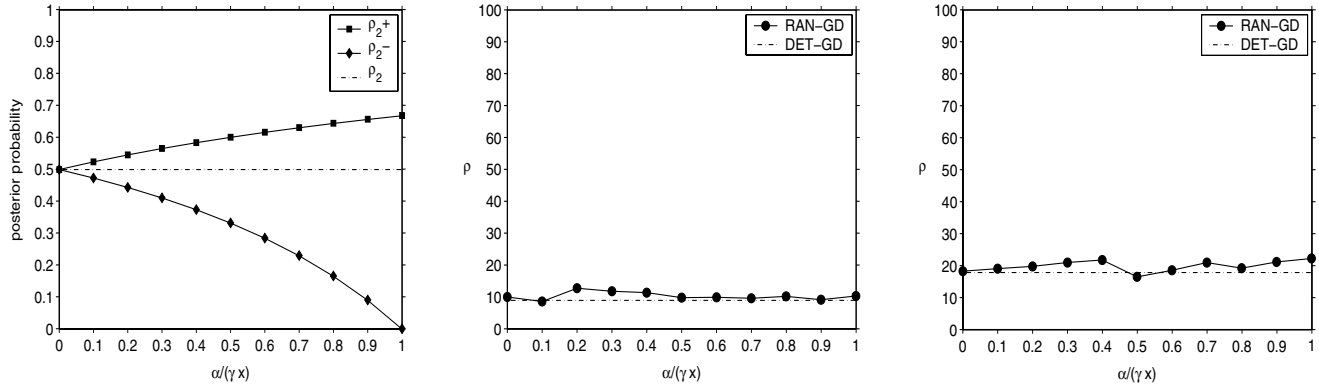
(a) (b) (c)

Figure 1. CENSUS: (a) Support error ρ (b) False negatives σ^- (c) False positives σ^+



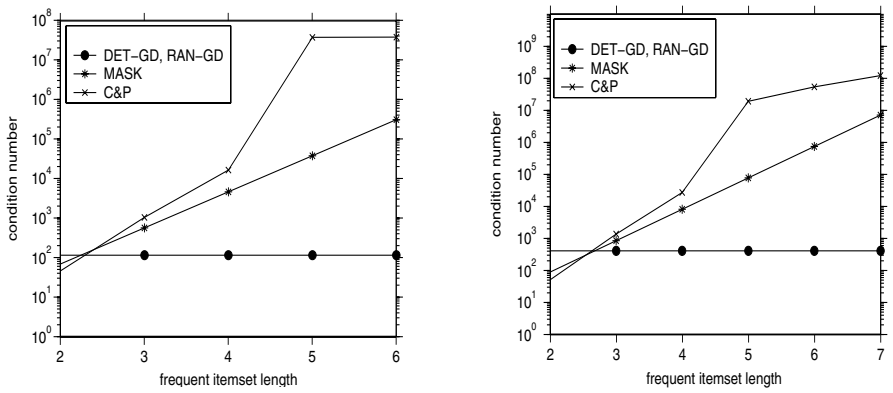
(a) (b) (c)

Figure 2. HEALTH: (a) Support error ρ (b) False negatives σ^- (c) False positives σ^+



(a) (b) (c)

Figure 3. (a) Posterior probability ranges (b) Support error ρ for CENSUS (c) Support error ρ for HEALTH



(a) (b)

Figure 4. Perturbation Matrix Condition Numbers: (a) CENSUS (b) HEALTH

diagonal-based techniques on the CENSUS and HEALTH datasets showed substantial reductions in frequent itemset identity and support reconstruction errors.

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

Acknowledgments

This work was supported in part by a Swarnajayanti Fellowship from the Dept. of Science & Technology, Govt. of India. We thank R. Vittal Rao and Md. Emtiyaz Khan for their technical inputs, and the referees for their constructive suggestions.

References

[1] C. Aggarwal and P. Yu. A condensation approach to privacy preserving data mining. *Proc. of 9th Intl. Conf. on Extending Database Technology (EDBT)*, March 2004.

[2] D. Agrawal and C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. *Proc. of ACM Symp. on Principles of Database Systems (PODS)*, May 2001.

[3] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *Proc. of ACM SIGMOD Intl. Conf. on Management of Data*, June 1993.

[4] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. *Proc. of 20th Intl. Conf. on Very Large Data Bases (VLDB)*, September 1994.

[5] R. Agrawal and R. Srikant. Privacy-preserving data mining. *Proc. of ACM SIGMOD Intl. Conf. on Management of Data*, May 2000.

[6] S. Agrawal and J. Haritsa. A framework for high-accuracy privacy-preserving mining. Technical Report TR-2004-02, Database Systems Lab, Indian Institute of Science, 2004. (http://dsl.serc.iisc.ernet.in/pub/TR/TR-2004-02.pdf).

[7] S. Agrawal and J. Haritsa. On addressing efficiency concerns in privacy-preserving mining. *Proc. of 9th Intl. Conf. on Database Systems for Advanced Applications (DASFAA)*, March 2004.

[8] P. de Wolf, J. Gouweleeuw, P. Kooiman, and L. Willenborg. Reflections on PRAM. *Proc. of Statistical Data Protection Conf.*, March 1998.

[9] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. *Proc. of ACM Symp. on Principles of Database Systems (PODS)*, June 2003.

[10] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. *Proc. of 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, July 2002.

[11] W. Feller. *An Introduction to Probability Theory and its Applications (Vol. I)*. Wiley, 1988.

[12] J. Gouweleeuw, P. Kooiman, L. Willenborg, and P. de Wolf. Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14(4), 1998.

[13] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. *Proc. of the 3rd IEEE Intl. Conf. on Data Mining (ICDM), Melbourne, Florida*, December 2003.

[14] S. Rizvi and J. Haritsa. Maintaining data privacy in association rule mining. *Proc. of 28th Intl. Conf. on Very Large Databases (VLDB)*, August 2002.

[15] G. Strang. *Linear Algebra and its Applications*. Thomson Learning Inc., 1988.

[16] Y. Wang. On the number of successes in independent trials. *Statistica Silica 3*, 1993.

[17] http://dataferrett.census.gov.

[18] http://www.ics.uci.edu/ mlearn/mlrepository.html.