Privacy Preserving Mining of Association Rules

Alexandre Evfimievski* Ramakrishnan Srikant Rakesh Agrawal Johannes Gehrke*

IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120, USA

ABSTRACT

We present a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions. While it is feasible to recover association rules and preserve privacy using a straightforward "uniform" randomization, the discovered rules can unfortunately be exploited to find privacy breaches. We analyze the nature of privacy breaches and propose a class of randomization operators that are much more effective than uniform randomization in limiting the breaches. We derive formulae for an unbiased support estimator and its variance, which allow us to recover itemset supports from randomized datasets, and show how to incorporate these formulae into mining algorithms. Finally, we present experimental results that validate the algorithm by applying it on real datasets.

1. INTRODUCTION

The explosive progress in networking, storage, and processor technologies is resulting in an unprecedented amount of digitization of information. It is estimated that the amount of information in the world is doubling every 20 months [20]. In concert with this dramatic and escalating increase in digital data, concerns about privacy of personal information have emerged globally [15] [17] [20] [24]. Privacy issues are further exacerbated now that the internet makes it easy for the new data to be automatically collected and added to databases [10] [13] [14] [27] [28] [29]. The concerns over massive collection of data are naturally extending to analytic tools applied to data. Data mining, with its promise to efficiently discover valuable, non-obvious information from large databases, is particularly vulnerable to misuse [11] [16] [20] [23].

An interesting new direction for data mining research is the development of techniques that incorporate privacy concerns [3]. The following question was raised in [7]: since the primary task in data mining is the development of models about aggregated data, can we develop accurate models without access to precise information in individual data records? Specifically, they studied the technical feasibility of building accurate classification models using training data in which the sensitive numeric values in a user's record have been randomized so that the true values cannot be estimated with sufficient precision. Randomization is done using the statistical method of value distortion [12] that returns a value $x_i + r$ instead of x_i where r is a random value drawn from some distribution. They proposed a Bayesian procedure for correcting perturbed distributions and presented three algorithms for building accurate decision trees [9] [21] that rely on reconstructed distributions.[1] In [2], the authors derived an Expectation Maximization (EM) algorithm for reconstructing distributions and proved that the EM algorithm converged to the maximum likelihood estimate of the original distribution based on the perturbed data. They also pointed out that the EM algorithm was in fact identical to the Bayesian reconstruction procedure in [7], except for an approximation (partitioning values into intervals) that was made by the latter.

1.1 Contributions of this Paper

We continue the investigation of the use of randomization in developing privacy-preserving data mining techniques, and extend this line of inquiry along two dimensions:

- categorical data instead of numerical data, and
- association rule mining [4] instead of classification.

We will focus on the task of finding frequent itemsets in association rule mining, which we briefly review next.

Definition 1. Suppose we have a set $\mathcal{I}$ of n items: $\mathcal{I} = \{a_1, a_2, \ldots, a_n\}$. Let T be a sequence of N transactions $T = (t_1, t_2, \ldots, t_N)$ where each transaction t_i is a subset of $\mathcal{I}$. Given an itemset $A \subset \mathcal{I}$, its *support* $\mathrm{supp}^T(A)$ is defined as

$$\mathrm{supp}^T(A) := \frac{\#\{t \in T \mid A \subseteq t\}}{N}. \tag{1}$$

An itemset $A \subset \mathcal{I}$ is called *frequent* in T if $\mathrm{supp}^T(A) \geqslant \tau$, where τ is a user-defined parameter.

We consider the following setting. Suppose we have a server and many clients. Each client has a set of items (e.g.,

*Department of Computer Science
Cornell University, Ithaca, NY 14853, USA

SIGKDD 02 Edmonton, Alberta, Canada
Copyright 2002 ACM 1-58113-567-X/02/0007 ...$5.00.

[1] Once we have reconstructed distributions, it is straightforward to build classifiers that assume independence between attributes, such as Naive Bayes [19] .

books or web pages or TV programs). The clients want the server to gather statistical information about associations among items, perhaps in order to provide recommendations to the clients. However, the clients do not want the server to know with certainty who has got which items. When a client sends its set of items to the server, it modifies the set according to some specific randomization policy. The server then gathers statistical information from the modified sets of items (transactions) and recovers from it the actual associations.

The following are the important results contained in this paper:

- In Section 2, we show that a straightforward uniform randomization leads to privacy breaches.
- We formally model and define privacy breaches in Section 3.
- We present a class of randomization operators in Section 4 that can be tuned for different tradeoffs between discoverability and privacy breaches. We derive formulae for the effect of randomization on support, and show how to recover the original support of an association from the randomized data.
- We present experimental results on two real datasets in Section 5, as well as graphs showing the relationship between discoverability, privacy, and data characteristics.

1.2 Related Work

There has been extensive research in the area of statistical databases motivated by the desire to provide statistical information (sum, count, average, maximum, minimum, pth percentile, etc.) without compromising sensitive information about individuals (see surveys in [1] [22].) The proposed techniques can be broadly classified into query restriction and data perturbation. The query restriction family includes restricting the size of query result, controlling the overlap amongst successive queries, keeping audit trail of all answered queries and constantly checking for possible compromise, suppression of data cells of small size, and clustering entities into mutually exclusive atomic populations. The perturbation family includes swapping values between records, replacing the original database by a sample from the same distribution, adding noise to the values in the database, adding noise to the results of a query, and sampling the result of a query. There are negative results showing that the proposed techniques cannot satisfy the conflicting objectives of providing high quality statistics and at the same time prevent exact or partial disclosure of individual information [1].

The most relevant work from the statistical database literature is the work by Warner [26], where he developed the "randomized response" method for survey results. The method deals with a single boolean attribute (e.g., drug addiction). The value of the attribute is retained with probability p and flipped with probability $1 - p$. Warner then derived equations for estimating the true value of queries such as COUNT (Age = 42 & Drug Addiction = Yes). The approach we present in Section 2 can be viewed as a generalization of Warner's idea.

Another related work is [25], where they consider the problem of mining association rules over data that is vertically partitioned across two sources, i.e, for each transaction, some of the items are in one source, and the rest in the other source. They use multi-party computation techniques for scalar products to be able to compute the support of an itemset (when the two subsets that together form the itemset are in different sources), without either source revealing exactly which transactions support a subset of the itemset. In contrast, we focus on preserving privacy when the data is horizontally partitioned, i.e., we want to preserve privacy for individual transactions, rather than between two data sources that each have a vertical slice.

Related, but not directly relevant to our current work, is the problem of inducing decision trees over horizontally partitioned training data originating from sources who do not trust each other. In [16], each source first builds a local decision tree over its true data, and then swaps values amongst records in a leaf node of the tree to generate randomized training data. Another approach, presented in [18], does not use randomization, but makes use of cryptographic oblivious functions during tree construction to preserve privacy of two data sources.

2. UNIFORM RANDOMIZATION

A straightforward approach for randomizing transactions would be to generalize Warner's "randomized response" method, described in Section 1.2. Before sending a transaction to the server, the client takes each item and with probability p replaces it by a new item not originally present in this transaction. Let us call this process *uniform* randomization.

Estimating true (nonrandomized) support of an itemset is nontrivial even for uniform randomization. Randomized support of, say, a 3-itemset depends not only on its true support, but also on the supports of its subsets. Indeed, it is much more likely that only one or two of the items are inserted by chance than all three. So, almost all "false" occurrences of the itemset are due to (and depend on) high subset supports. This requires estimating the supports of all subsets simultaneously. (The algorithm is similar to the algorithm presented in Section 4 for select-a-size randomization, and the formulae from Statements 1, 3 and 4 apply here as well.) For large values of p, most of the items in most randomized transactions will be "false", so we seem to have obtained a reasonable privacy protection. Also, if there are enough clients and transactions, then frequent itemsets will still be "visible", though less frequent than originally. For instance, after uniform randomization with $p = 80\%$, an itemset of 3 items that originally occurred in 1% transactions will occur in about $1\% \cdot (0.2)^3 = 0.008\%$ transactions, which is about 80 transactions per each million. The opposite effect of "false" itemsets becoming more frequent is comparatively negligible if there are many possible items: for 10,000 items, the probability that, say, 10 randomly inserted items contain a given 3-itemset is less than $10^{-7}\%$.

Unfortunately, this randomization has a problem. If we know that our 3-itemset escapes randomization in 80 per million transactions, and that it is unlikely to occur even once *because of* randomization, then every time we see it in a randomized transaction we know with near certainty of its presence in the nonrandomized transaction. With even more certainty we will know that at least one item from this itemset is "true": as we have mentioned, a chance insertion of only one or two of the items is much more likely than of all three. In this case we can say that a *privacy breach* has occurred. Although privacy is preserved on average, personal information leaks through uniform randomization for some fraction of transactions, despite the high value of p.

The rest of the paper is devoted to defining a framework for studying privacy breaches and developing techniques for finding frequent itemsets while avoiding breaches.

3. PRIVACY BREACHES

Definition 2. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space of elementary events over some set Ω and σ-algebra $\mathcal{F}$. A *randomization operator* is a measurable function

$$R : \Omega \times \{\text{all possible } T\} \to \{\text{all possible } T\}$$

that randomly transforms a sequence of N transactions into a (usually) different sequence of N transactions. Given a sequence of N transactions T, we shall write $T' = R(T)$, where T is constant and $R(T)$ is a random variable.

Definition 3. Suppose that a nonrandomized sequence T is drawn from some known distribution, and $t_i \in T$ is the i-th transaction in T. A *general privacy breach of level ρ* with respect to a property $P(t_i)$ occurs if

$$\exists T' : \ \mathbf{P}\left[P(t_i) \mid R(T) = T'\right] \geqslant \rho.$$

We say that a property $Q(T')$ *causes a privacy breach* of level ρ with respect to $P(t_i)$ if

$$\mathbf{P}\left[P(t_i) \mid Q(R(T))\right] \geqslant \rho.$$

When we define privacy breaches, we think of the prior distribution of transactions as known, so that it makes sense to speak about a posterior probability of a property $P(t_i)$ versus prior. In practice, however, we do not know the prior distribution. In fact, there is no prior distribution; the transactions are not randomly generated. However, modeling transactions as being randomly generated from a prior distribution allows us to cleanly define privacy breaches.

Consider a situation when, for some transaction $t_i \in T$, an itemset $A \subseteq \mathcal{I}$ and an item $a \in A$, the property "$A \subseteq t'_i \in T'$" causes a privacy breach w. r. t. the property "$a \in t_i$." In other words, the presence of A in a randomized transaction makes it likely that item a is present in the corresponding nonrandomized transaction.

Definition 4. We say that itemset A *causes a privacy breach of level ρ* if for some item $a \in A$ and some $i \in 1 \ldots N$ we have $\mathbf{P}\left[a \in t_i \mid A \subseteq t'_i\right] \geqslant \rho$.

We will focus on controlling the class of privacy breaches given by Definition 4. Thus we ignore the effect of other information the server obtains from a randomized transaction, such as which items the randomized transaction does not contain, or the randomized transaction size. We also do not attempt to control breaches that occur because the server knows some other information about items and clients besides the transactions. For example, the server may know some geographical or demographic data about the clients. Finally, in Definition 4, we only considered positive breaches, i.e., we know with high probability that an item was present in the original transaction. In some scenarios, being confident that an item was *not* present in the original transaction may also be considered a privacy breach.

4. ALGORITHM

> "Where does a wise man hide a leaf? In the forest. But what does he do if there is no forest?" ... "He grows a forest to hide it in." – G.K. Chesterton, "The Sign of the Broken Sword"

The intuition of breach control is quite simple: in addition to replacing some of the items, we shall insert so many "false" items into a transaction that one is as likely to see a "false" itemset as a "true" one.

4.1 Randomization Operators

Definition 5. We call randomization R a *per-transaction* randomization if, for $T = (t_1, t_2, \ldots, t_N)$, we can represent $R(T)$ as

$$R(t_1, t_2, \ldots, t_N) = (R(1, t_1), R(2, t_2), \ldots, R(N, t_N)),$$

where $R(i, t)$ are independent random variables whose distributions depend only on t (and not on i). We shall write $t'_i = R(i, t_i) = R(t_i)$.

Definition 6. A randomization operator R is called *item-invariant* if, for every transaction sequence T and for every permutation $\pi : \mathcal{I} \to \mathcal{I}$ of items, the distribution of $\pi^{-1} R(\pi T)$ is the same as of $R(T)$. Here πT means the application of π to all items in all transactions of T at once.

Definition 7. A *select-a-size* randomization operator has the following parameters, for each possible input transaction size m:

- Default probability of an item (also called *randomization level*) $\rho_m \in (0, 1)$;
- Transaction subset size selection probabilities $p_m[0], p_m[1], \ldots, p_m[m]$, such that every $p_m[j] \geqslant 0$ and

$$p_m[0] + p_m[1] + \ldots + p_m[m] = 1.$$

Given a sequence of transactions $T = (t_1, t_2, \ldots, t_N)$, the operator takes each transaction t_i independently and proceeds as follows to obtain transaction t'_i ($m = |t_i|$).

1. The operator selects an integer j at random from the set $\{0, 1, \ldots, m\}$ so that $\mathbf{P}\left[j \text{ is selected}\right] = p_m[j]$.
2. It selects j items from t_i, uniformly at random (without replacement). These items, and no other items of t_i, are placed into t'_i.
3. It considers each item $a \notin t_i$ in turn and tosses a coin with probability ρ_m of "heads" and $1 - \rho_m$ of "tails". All those items for which the coin faces "heads" are added to t'_i.

Remark 1. Both uniform (Section 2) and select-a-size operators are per-transaction because they apply the same randomization algorithm to each transaction independently. They are also item-invariant since they do not use any item-specific information (if we rename or reorder the items, the outcome probabilities will not be affected).

Definition 8. A *cut-and-paste* randomization operator is a special case of a select-a-size operator (and which we shall actually test on datasets). For each possible input transaction size m, it has two parameters: $\rho_m \in (0,1)$ (randomization level) and an integer $K_m > 0$ (the *cutoff*). The operator takes each input transaction t_i independently and proceeds as follows to obtain transaction t'_i (here $m = |t_i|$) :

1. It chooses an integer j uniformly at random between 0 and K_m; if $j > m$, it sets $j = m$.
2. The operator selects j items out of t_i uniformly at random (without replacement). These items are placed into t'_i.
3. Each other item (including the rest of t_i) is placed into t'_i with probability ρ_m, independently.

Remark 2. For any m, a cut-and-paste operator has only two parameters, ρ_m and K_m, to play with; moreover, K_m is an integer. Because it is easy to find optimal values for these parameters (Section 4.4), we chose to test this operator, leaving open the problem of optimizing the m parameters of the "unabridged" select-a-size. To see that cut-and-paste is a case of select-a-size, let us write down the formulae for the $p_m[j]$'s:

$$p_m[j] = \sum_{i=0}^{\min\{K,j\}} \binom{m-i}{j-i} \rho^{j-i}(1-\rho)^{m-j} \cdot \cdot \begin{cases} 1 - m/(K+1) & \text{if } i = m \text{ and } i < K \\ 1/(K+1) & \text{otherwise} \end{cases}$$

Now let us give one example of a randomization operator that is not a per-transaction randomization, because it uses the knowledge of several transactions per each randomized transaction.

Example 1. The *mixing* randomization operator has one integer parameter $K \geqslant 2$ and one real-valued parameter $p \in (0,1)$. Given a sequence of transactions $T = (t_1, t_2, \ldots, t_N)$, the operator takes each transaction t_i independently and proceeds as follows to obtain transaction t'_i:

1. Other than t_i, pick $K-1$ more transactions (with replacement) from T and union the K transactions as sets of items. Let t''_i be this union.
2. Consider each item $a \in t''_i$ in turn and toss a coin with probability p of "heads" and $1-p$ of "tails".
3. All those items for which the coin faces "tails" are removed from the transaction. The remaining items constitute the randomized transaction.

For the purpose of privacy-preserving data mining, it is natural to focus mostly on per-transaction randomizations, since they are the easiest and safest to implement. Indeed, a per-transaction randomization does not require the users (who submit randomized transactions to the server) to communicate with each other in any way, nor to exchange random bits. On the contrary, implementing mixing randomization, for example, requires to organize an exchange of nonrandomized transactions between users, which opens an opportunity for cheating or eavesdropping.

4.2 Effect of Randomization on Support

Let T be a sequence of transactions of length N, and let A be some subset of items (that is, $A \subseteq \mathcal{I}$). Suppose we randomize T and get $T' = R(T)$. The support $s' = \operatorname{supp}^{T'}(A)$ of A for T' is a random variable that depends on the outcome of randomization. Here we are going to determine the distribution of s', under the assumption of having a per-transaction and item-invariant randomization.

Definition 9. The fraction of the transactions in T that have intersection with A of size l among all transactions in T is called *partial support* of A for intersection size l:

$$\operatorname{supp}_l^T(A) := \frac{\#\{t \in T \mid \#(A \cap t) = l\}}{N}. \qquad (2)$$

It is easy to see that $\operatorname{supp}^T(A) = \operatorname{supp}_k^T(A)$ for $k = |A|$, and that

$$\sum_{l=0}^{k} \operatorname{supp}_l^T(A) = 1$$

since those transactions in T that do not intersect A at all are covered in $\operatorname{supp}_0^T(A)$.

Definition 10. Suppose that our randomization operator is both per-transaction and item-invariant. Consider a transaction t of size m and an itemset $A \subset \mathcal{I}$ of size k. After randomization, transaction t becomes t'. We define

$$p_k^m\,[l \to l'] = p\,[l \to l'] := \mathbf{P}\,[\#(t' \cap A) = l' \mid \#(t \cap A) = l]. \qquad (3)$$

Here both l and l' must be integers in $\{0, 1, \ldots, k\}$.

Remark 3. The value of $p_k^m\,[l \to l']$ is well-defined (does not depend on any other information about t and A, or other transactions in T and T' besides t and t'). Indeed, because we have a per-transaction randomization, the distribution of t' depends neither on other transactions in T besides t, nor on their randomized outcomes. If there were other t_1 and B with the same (m, k, l), but a different probability (3) for the same l', we could consider a permutation π of $\mathcal{I}$ such that $\pi t = t_1$ and $\pi A = B$; the application of π or of π^{-1} would preserve intersection sizes l and l'. By item-invariance we have

$$\mathbf{P}\,[\#(t' \cap A) = l'] = \mathbf{P}\,[\#(\pi^{-1}R(\pi t)\, \cap\, A) = l'],$$

but by the choice of π we also have

$$\mathbf{P}\,[\#(\pi^{-1}R(\pi t)\, \cap\, A) = l'] = \mathbf{P}\,[\#(\pi^{-1}R(t_1)\, \cap\, \pi^{-1}B) = l'] = \mathbf{P}\,[\#(t'_1 \cap B) = l'] \neq \mathbf{P}\,[\#(t' \cap A) = l'],$$

a contradiction.

STATEMENT 1. *Suppose that our randomization operator is both per-transaction and item-invariant. Suppose also that all the N transactions in T have the same size m. Then, for a given subset $A \subseteq \mathcal{I}$, $|A| = k$, the random vector*

$$N \cdot (s'_0, s'_1, \ldots, s'_k), \quad \textit{where } s'_l := \operatorname{supp}_l^{T'}(A) \qquad (4)$$

is a sum of $k+1$ independent random vectors, each having a multinomial distribution. Its expected value is given by

$$\mathbf{E}\,(s'_0, s'_1, \ldots, s'_k)^T = P \cdot (s_0, s_1, \ldots, s_k)^T \qquad (5)$$

where P is the $(k+1)\times(k+1)$ matrix with elements $P_{l'\,l} = p\,[l \to l']$, and the covariance matrix is given by

$$\mathbf{Cov}\,(s'_0, s'_1, \ldots, s'_k)^T = \frac{1}{N}\cdot\sum_{l=0}^{k} s_l\, D[l] \qquad (6)$$

where each $D[l]$ is a $(k+1)\times(k+1)$ matrix with elements

$$D[l]_{i\,j} = p\,[l \to i]\cdot\delta_{i=j} - p\,[l \to i]\cdot p\,[l \to j]. \qquad (7)$$

Here s_l denotes $\text{supp}^T_l(A)$, and the T over vectors denotes the transpose operation; $\delta_{i=j}$ is one if $i = j$ and zero otherwise.

Proof. See Appendix A.1. □

Remark 4. In Statement 1 we have assumed that all transactions in T have the same size. If this is not so, we have to consider each transaction size separately and then use per-transaction independence.

Statement 2. *For a select-a-size randomization with randomization level ρ and size selection probabilities $\{p_m[j]\}$, we have:*

$$p_k^m\,[l \to l'] \;=\; \sum_{j=0}^{m} p_m[j]\cdot \sum_{q\,=\,\max\{0,\,j+l-m,\,l+l'-k\}}^{\min\{j,l,l'\}} \frac{\binom{l}{q}\binom{m-l}{j-q}}{\binom{m}{j}}\cdot$$

$$\cdot\binom{k-l}{l'-q}\,\rho^{l'-q}(1-\rho)^{k-l-l'+q}. \qquad (8)$$

Proof. See Appendix A.2. □

4.3 Support Recovery

Let us assume that all transactions in T have the same size m, and let us denote

$$\vec{s} := (s_0, s_1, \ldots, s_k)^T, \quad \vec{s}\,' := (s_0, s_1, \ldots, s_k)^T;$$

then, according to (5), we have

$$\mathbf{E}\;\vec{s}\,' = P\cdot\vec{s}. \qquad (9)$$

Denote $Q = P^{-1}$ (assume that it exists) and multiply both sides of (9) by Q:

$$\vec{s} = Q\cdot\mathbf{E}\;\vec{s}\,' = \mathbf{E}\;Q\cdot\vec{s}\,'.$$

We have thus obtained an unbiased estimator for the original partial supports given randomized partial supports:

$$\vec{s}_{\text{est}} := Q\cdot\vec{s}\,' \qquad (10)$$

Using (6), we can compute the covariance matrix of $\vec{s}_{\text{est}}$:

$$\mathbf{Cov}\;\vec{s}_{\text{est}} = \mathbf{Cov}\,(Q\cdot\vec{s}\,') = Q\,(\mathbf{Cov}\;\vec{s}\,')\,Q^T =$$

$$= \frac{1}{N}\cdot\sum_{l=0}^{k} s_l\,Q\,D[l]\,Q^T. \qquad (11)$$

If we want to estimate this covariance matrix by looking only at randomized data, we may use $\vec{s}_{\text{est}}$ instead of $\vec{s}$ in (11):

$$(\mathbf{Cov}\;\vec{s}_{\text{est}})_{\text{est}} = \frac{1}{N}\cdot\sum_{l=0}^{k}(\vec{s}_{\text{est}})_l\,Q\,D[l]\,Q^T.$$

This estimator is also unbiased:

$$\mathbf{E}\,(\mathbf{Cov}\;\vec{s}_{\text{est}})_{\text{est}} = \frac{1}{N}\cdot\sum_{l=0}^{k}(\mathbf{E}\;\vec{s}_{\text{est}})_l\,Q\,D[l]\,Q^T = \mathbf{Cov}\;\vec{s}_{\text{est}}.$$

In practice, we want only the k-th coordinate of $\vec{s}$, that is, the support $s = \text{supp}^T(A)$ of our itemset A in T. We denote by $\tilde{s}$ the k-th coordinate of $\vec{s}_{\text{est}}$, and use $\tilde{s}$ to estimate s. Let us compute simple formulae for $\tilde{s}$, its variance and the unbiased estimator of its variance. Denote

$$q\,[l \leftarrow l'] := Q_{l\,l'}.$$

Statement 3.

$$\tilde{s} = \sum_{l'=0}^{k} s'_{l'}\cdot q\,[k \leftarrow l']\,;$$

$$\mathbf{Var}\;\tilde{s} = \frac{1}{N}\sum_{l=0}^{k} s_l\,\Big(\sum_{l'=0}^{k} p\,[l \to l']\;q\,[k \leftarrow l']^2 - \delta_{l=k}\Big);$$

$$(\mathbf{Var}\;\tilde{s})_{est} = \frac{1}{N}\sum_{l'=0}^{k} s'_{l'}\,\big(q\,[k \leftarrow l']^2 - q\,[k \leftarrow l']\,\big).$$

Proof. See Appendix A.3. □

We conclude this subsection by giving a linear coordinate transformation in which the matrix P from Statement 1 becomes triangular. (We use this transformation for privacy breach analysis in Section 4.4.) The coordinates after the transformation have a combinatorial meaning, as given in the following definition.

Definition 11. Suppose we have a transaction sequence T and an itemset $A \subseteq \mathcal{I}$. Given an integer l between 0 and $k = |A|$, consider all subsets $C \subseteq A$ of size l. The sum of supports of all these subsets is called the *cumulative support* for A of order l and is denoted as follows:

$$\Sigma_l \;=\; \Sigma_l(A,T) \;:=\; \sum_{C\subseteq A,\;|C|\,=\,l} \text{supp}^T(C)\,,$$

$$\vec{\Sigma} \;:=\; (\Sigma_0, \Sigma_1, \ldots, \Sigma_k)^T \qquad (12)$$

Statement 4. *The vector $\vec{\Sigma}$ of cumulative supports is a linear transformation of the vector $\vec{s}$ of partial supports, namely,*

$$\Sigma_l \;=\; \sum_{j=l}^{k}\binom{j}{l}s_j \quad \textit{and} \quad s_l \;=\; \sum_{j=l}^{k}(-1)^{j-l}\binom{j}{l}\Sigma_j; \qquad (13)$$

in the $\vec{\Sigma}$ and $\vec{\Sigma}\,'$ space (instead of $\vec{s}$ and $\vec{s}\,'$) matrix P is lower triangular.

Proof. See Appendix A.4. □

4.4 Limiting Privacy Breaches

Here we determine how privacy depends on randomization. We shall use Definition 4 and assume a per-transaction and item-invariant randomization.

Consider some itemset $A \subseteq \mathcal{I}$ and some item $a \in A$; fix a transaction size m. We shall assume that m is known to the server, so that we do not have to combine probabilities for different nonrandomized sizes. Assume also that a partial support $s_l = \mathrm{supp}_l^T(A)$ approximates the corresponding prior probability $\mathbf{P}\,[\#(t \cap A) = l]$. Suppose we know the following prior probabilities:

$$\begin{aligned} s_l^+ &:= \mathbf{P}\,[\#(t \cap A) = l,\ a \in t], \\ s_l^- &:= \mathbf{P}\,[\#(t \cap A) = l,\ a \notin t]. \end{aligned}$$

Notice that $s_l = s_l^+ + s_l^-$ simply because

$$\#(t \cap A) = l \;\Leftrightarrow\; \left[\begin{array}{l} a \in t \;\&\; \#(t \cap A) = l, \text{ or} \\ a \notin t \;\&\; \#(t \cap A) = l. \end{array}\right.$$

Let us use these priors and compute the posterior probability of $a \in t$ given $A \subseteq t'$:

$$\begin{aligned} \mathbf{P}\,[a \in t \mid A \subseteq t'] \;&=\; \frac{\mathbf{P}\,[a \in t,\ A \subseteq t']}{\mathbf{P}\,[A \subseteq t']} \;= \\ = \sum_{l=1}^{k} \mathbf{P}\,[\#(t \cap A) = l,\ a \in t,\ A \subseteq t'] &\Big/ \sum_{l=0}^{k} s_l \cdot p\,[l \to k] \\ = \sum_{l=1}^{k} \mathbf{P}\,[\#(t \cap A) = l,\ a \in t] \cdot p\,[l \to k] &\Big/ \sum_{l=0}^{k} s_l \cdot p\,[l \to k] \\ = \sum_{l=1}^{k} s_l^+ \cdot p\,[l \to k] &\Big/ \sum_{l=0}^{k} s_l \cdot p\,[l \to k]. \end{aligned}$$

Thus, in order to prevent privacy breaches of level 50% as defined in Definition 4, we need to ensure that always

$$\sum_{l=1}^{k} s_l^+ \cdot p\,[l \to k] \;<\; 0.5 \cdot \sum_{l=0}^{k} s_l \cdot p\,[l \to k]\,. \qquad (14)$$

The problem is that we have to randomize the data *before* we know any supports. Also, we may not have the luxury of setting "oversafe" randomization parameters because then we may not have enough data to perform a reasonably accurate support recovery. One way to achieve a compromise is to:

1. Estimate maximum possible support $s_{max}(k, m)$ of a k-itemset in the transactions of given size m, for different k and m;

2. Given the maximum supports, find values for s_l and s_l^+ that are most likely to cause a privacy breach;

3. Make randomization just strong enough to prevent such a privacy breach.

Since $s_0^+ = 0$, the most privacy-challenging situations occur when s_0 is small, that is, when our itemset A and its subsets are frequent.

In our experiments we consider a privacy-challenging k-itemset A such that, for every $l > 0$, all its subsets of size l have the maximum possible support $s_{max}(l, m)$. The partial supports for such a test-itemset are computed from the cumulative supports Σ_l using Statement 4. By it and by (12), we have ($l > 0$)

$$s_l = \sum_{j=l}^{k} (-1)^{j-l} \binom{j}{l} \Sigma_j, \quad \Sigma_j = \binom{k}{j} s_{max}(j, m) \qquad (15)$$

since there are $\binom{k}{j}$ j-subsets in A. The values of s_l^+ follow if we note that all l-subsets of A, with a and without, appear equally frequently as $t \cap A$:

$$\begin{aligned} s_l^+ \;:=\; &\mathbf{P}\,[\#(t \cap A) = l,\ a \in t] \;= \\ &= \mathbf{P}\,[a \in t \mid \#(t \cap A) = l] \cdot s_l \;=\; l/k \cdot s_l. \qquad (16) \end{aligned}$$

While one can construct cases that are even more privacy-challenging (for example, if $a \in A$ occurs in a transaction every time any nonempty subset of A does), we found the above model (15) and (16) to be sufficiently pessimistic on our datasets.

We can now use these formulae to obtain cut-and-paste randomization parameters ρ_m and K_m as follows. Given m, consider all cutoffs from $K_m = 3$ to some $K_{\max}$ (usually this $K_{\max}$ equals the maximum transaction size) and determine the smallest randomization levels $\rho_m(K_m)$ that satisfy (14). Then select (K_m, ρ_m) that gives the best discoverability (by computing the lowest discoverable supports, see Section 5.1).

4.5 Discovering Associations

We show how to discover itemsets with high true support given a set of randomized transactions. Although we use the Apriori algorithm [5] to make the ideas concrete, the modifications directly apply to any algorithm that uses Apriori candidate generation, i.e., to most current association discovery algorithms.[2] The key *lattice property* of supports used by *Apriori* is that, for any two itemsets $A \subseteq B$, the true support of A is equal to or larger than the true support of B. A simplified version of *Apriori*, given a (nonrandomized) transactions file and a minimum support $s_{\min}$, works as follows:

1. Let $k = 1$, let "candidate sets" be all single items. Repeat the following until no candidate sets are left:
   (a) Read the data file and compute the supports of all candidate sets;
   (b) Discard all candidate sets whose support is below $s_{\min}$;
   (c) Save the remaining candidate sets for output;
   (d) Form all possible $(k+1)$-itemsets such that all their k-subsets are among the remaining candidates. Let these itemsets be the new candidate sets.
   (e) Let $k = k + 1$.

2. Output all the saved itemsets.

It is (conceptually) straightforward to modify this algorithm so that now it reads the randomized dataset, computes partial supports of all candidate sets (for all nonrandomized transaction sizes) and recovers their predicted supports and sigmas using the formulae from Statement 3. However, for the predicted supports the lattice property is no longer true. It is quite likely that for an itemset that is slightly above minimum support and whose predicted support is also above minimum support, that one of its subsets will have predicted support below minimum support. So if we discard all candidates below minimum support for the purpose of candidate generation, we will miss many (perhaps even the majority)

[2] The main class of algorithms where this would not apply are those that find only maximal frequent itemsets, e.g., [8]. However, randomization precludes finding very long itemsets, so this is a moot point.

of the longer frequent itemsets. Hence, for candidate generation, we discard only those candidates whose predicted support is "significantly" smaller than $s_{\min}$, where significance is measured by means of predicted sigmas. Here is the modified version of *Apriori*:

```
1. Let $k = 1$, let "candidate sets" be all single-item sets. Repeat the following until $k$ is too large for support recovery (or until no candidate sets are left):
    (a) Read the randomized data file and compute the partial supports of all candidate sets, separately for each nonrandomized transaction size[3];
    (b) Recover the predicted supports and sigmas for the candidate sets;
    (c) Discard every candidate set whose support is below its *candidate limit*;
    (d) Save for output only those candidate sets whose predicted support is at least $s_{\min}$;
    (e) Form all possible $(k + 1)$-itemsets such that all their $k$-subsets are among the remaining candidates. Let these itemsets be the new candidate sets.
    (f) Let $k = k + 1$.
2. Output all the saved itemsets.
```

We tried $s_{\min} - \sigma$ and $s_{\min} - 2\sigma$ as the candidate limit, and found that the former does a little better than the latter. It prunes more itemsets and therefore makes the algorithm work faster, and, when it discards a subset of an itemset with high predicted support, it usually turns out that the true support of this itemset is not as high.

5. EXPERIMENTAL RESULTS

Before we come to the experiments with datasets, we first show in Section 5.1 how our ability to recover supports depends on the permitted breach level, as well as other data characteristics. We then describe the real-life datasets in Section 5.2, and present results on these datasets in Section 5.3.

5.1 Privacy, Discoverability and Dataset Characteristics

We define the *lowest discoverable support* as the support at which the predicted support of an itemset is four sigmas away from zero, i.e, we can clearly distinguish the support of this itemset from zero. In practice, we may achieve reasonably good results even if the minimum support level is slightly lower than four sigma (as was the case for 3-itemsets in the randomized `soccer`, see below). However, the lowest discoverable support is a nice way to illustrate the interaction between discoverability, privacy breach levels, and data characteristics.

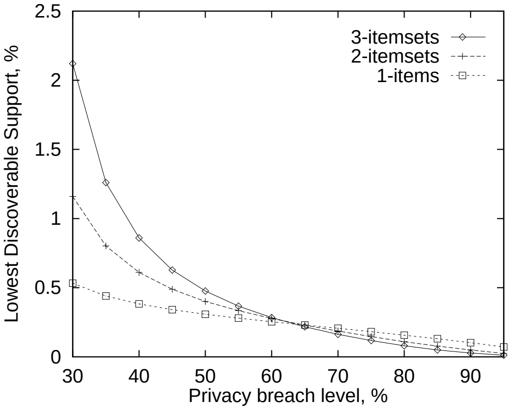

Figure 1: Lowest discoverable support for different breach levels. Transaction size is 5, five million transactions.

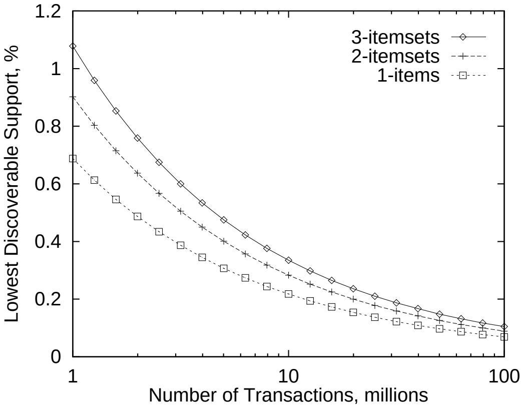

Figure 2: Lowest discoverable support versus number of transactions. Transaction size is 5, breach level is 50%.

Figure 1 shows how the lowest discoverable support changes with the privacy breach level. For higher privacy breach levels such as 95% (which could be considered a "plausible denial" breach level), we can discover 3-itemsets at very low supports. For more conservative privacy breach levels such as 50%, the lowest discoverable support is significantly higher. It is interesting to note that at higher breach levels (i.e. weaker randomization) it gets harder to discover 1-itemset supports than 3-itemset supports. This happens because the variance of a 3-itemset predictor depends highly nonlinearly on the amount of false items added while randomizing. When we add fewer false items at higher breach levels, we generate so much fewer false 3-itemset positives than false 1-itemset positives that 3-itemsets get an advantage over single items.

Figure 2 shows that the lowest discoverable support is roughly inversely proportional to the square root of the number of transactions. Indeed, the lowest discoverable support is defined to be proportional to the standard deviation (square root of the variance) of this support's prediction. If all the partial supports are fixed, the prediction's variance is inversely proportional to the number N of transactions according to Statement 3. In our case, the partial supports depend on N (because the lowest discoverable support does), i.e. they are not fixed; however, this does not appear to affect the variance very significantly (but justifies the word "roughly").

Finally, Figure 3 shows that transaction size has a sig-

[3] In our experiments, the nonrandomized transaction size is always known and included as a field into every randomized transaction

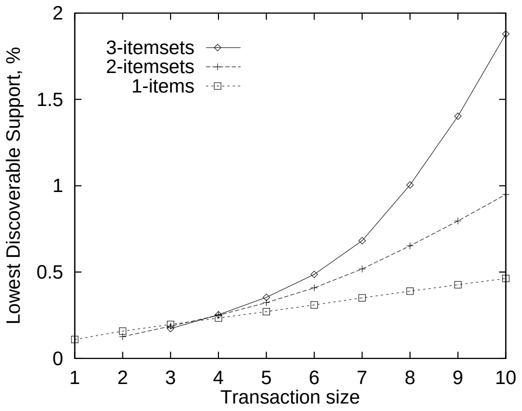

Figure 3: Lowest discoverable support for different transaction sizes. Five million transactions, breach level is 50%.

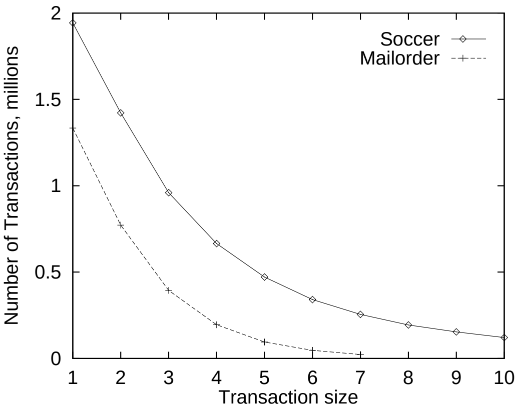

Figure 4: Number of transactions for each transaction size in the soccer and mailorder datasets.

nificant influence on support discoverability. In fact, for transactions of size 10 and longer, it is typically not possible to make them both breach-safe and simultaneously get useful information for mining transactions. Intuitively, a long transaction contains too much personal information to hide, because it may contain long frequent itemsets whose appearance in the randomized transaction could result in a privacy breach. We have to insert a lot of false items and cut off many true ones to ensure that such a long itemset in the randomized transaction is about as likely to be a false positive as to be a true positive. Such a strong randomization causes an exceedingly high variance in the support predictor for 2- and especially 3-itemsets, since it drives down their probability to "tunnel" through while raising high the probability of a false positive. In both our datasets we discard long transactions. The question of how to safely randomize and mine long transactions is left open.

5.2 The Datasets

We experimented with two "real-life" datasets. The `soccer` dataset is generated from the clickstream log of the 1998 World Cup Web site, which is publicly available at **ftp://researchsmp2.cc.vt.edu/pub/worldcup/**[4]. We scanned the log and produced a transaction file, where each transaction is a session of access to the site by a client. Each item in the transaction is a web request. Not all web requests were turned into items; to become an item, the request must satisfy the following:

1. Client's request method is GET;
2. Request status is OK;
3. File type is HTML.

A session starts with a request that satisfies the above properties, and ends when the last click from this ClientID timeouts. The timeout is set as 30 minutes. All requests in a session have the same ClientID. The `soccer` transaction file was then processed further: we deleted from all transactions the items corresponding to the French and English front page frames, and then we deleted all empty transactions and all transactions of size above 10. The resulting `soccer` dataset consists of 6, 525, 879 transactions, distributed as shown in Fig. 4.

The `mailorder` dataset is the same as that used in [6]. The original dataset consisted of around 2.9 million transactions, 15,836 items, and around 2.62 items per transaction. Each transaction was the set of items purchased in a single mail order. However, very few itemsets had reasonably high supports. For instance, there were only two 2-itemsets with support $\geqslant 0.2\%$, only five 3-itemsets with support $\geqslant 0.05\%$. Hence we decided to substitute all items by their parents in the taxonomy, which had reduced the number of items from 15836 to 96. It seems that, in general, moving items up the taxonomy is a natural thing to do for preserving privacy without losing aggregate information. We also discarded all transactions of size $\geqslant 8$ (which was less than 1% of all transactions) and finally obtained a dataset containing 2, 859, 314 transactions (Fig. 4).

5.3 The Results

We report the results for both datasets at a minimum support that is close to the lowest discoverable support, in order to show the resilience of our algorithm even at these very low support levels. We targeted a conservative breach level of 50%, so that, given a randomized transaction, for any item in the transaction it is at least as likely that someone did not buy that item (or access a web page) as that they did buy that item.

We used cut-and-paste randomization (see Definition 8) that has only two parameters, randomization level and cutoff, per each transaction size. We chose a cutoff of 7 for our experiments as a good compromise between privacy and discoverability. Given the values of maximum supports, we then used the methodology from Section 4.4 to find the lowest randomization level such that the breach probability (for each itemset size) is still below the desired breach level. The actual parameters (K_m is the cutoff, ρ_m is the randomization level for transaction size m) for `soccer` were:

m	1	2	3	4	5	6	7	8	9	10
K_m	7	7	7	7	7	7	7	7	7	7
ρ_m%	4.7	16.8	21.4	32.2	35.3	42.9	46.1	42.0	40.9	39.5

and for `mailorder` were:

m	1	2	3	4	5	6	7
K_m	7	7	7	7	7	7	7
ρ_m%	8.9	20.4	25.0	33.4	43.5	50.5	59.2

[4] M. Arlitt and T. Jin, "1998 World Cup Web Site Access Logs", August 1998. Available at **http://www.acm.org/sigcomm/ITA/**

Table 1 shows what happens if we mine itemsets from both randomized and nonrandomized files and then compare the results. We can see that, even for a low minimum support of 0.2%, most of the itemsets are mined correctly from the randomized file. There are comparatively few false positives (itemsets wrongly included into the output) and even fewer false drops (itemsets wrongly omitted). The predicted sigma for 3-itemsets ranges in 0.066−0.07% for `soccer` and in 0.047−0.048% for `mailorder`; for 2- and 1-itemsets sigmas are even less.

One might be concerned about the true supports of the false positives. Since we know that there are *many* more low-supported itemsets than there are highly supported, we might wonder whether most of the false positives are outliers, that is, have true support near zero. We have indeed seen outliers; however, it turns out that most of the false positives are not so far off. The tables 2 and 3 show that usually the true supports of false positives, as well as the predicted supports of false drops, are closer to 0.2% than to zero. This good news demonstrates the promise of randomization as a practical privacy-preserving approach.

Privacy Analysis We evaluate privacy breaches, i.e., the conditional probabilities from Definition 4, as follows. We count the occurrences of an itemset in a randomized transaction and its sub-items in the corresponding nonrandomized transaction. For example, assume an itemset $\{a, b, c\}$ occurs 100 times in the randomized data among transactions of length 5. Out of these 100 occurrences, 60 of the corresponding original transactions had the item b. We then say that this itemset caused a 60% privacy breach for transactions of length 5, since for these 100 randomized transactions, we estimate with 60% confidence that the item b was present in the original transaction.

Out of all sub-items of an itemset, we choose the item that causes the worst privacy breach. Then, for each combination of transaction size and itemset size, we compute over all frequent[5] itemsets the worst and the average value of this breach level. Finally, we pick the itemset size that gave the worst value for each of these two values.

Table 4 shows the results of the above analysis. To the left of the semicolon is the itemset size that was the worst. For instance, for all transactions of length 5 for soccer, the worst average breach was with 4-itemsets (43.9% breach), and the worst breach was with a 5-itemset (49.7% breach). We can see that, apart from fluctuations, the 50% level is observed everywhere except of a little "slip" for 9- and 10-item transactions of `soccer`. The "slip" resulted from our decision to use the corresponding maximal support information only for itemset sizes up to 7 (while computing randomization parameters).[6] However, since such long associations cannot be discovered, in practice, we will not get privacy breaches above 50%.

Summary Despite choosing a conservative privacy breach level of 50%, and further choosing a minimum support around the lowest discoverable support, we were able to successfully find most of the frequent itemsets, with relatively small numbers of false drops and false positives.

(a) mailorder, 0.2% minimum support

Itemset Size	True Itemsets	True Positives	False Drops	False Positives
1	65	65	0	0
2	228	212	16	28
3	22	18	4	5

(b) soccer, 0.2% minimum support

Itemset Size	True Itemsets	True Positives	False Drops	False Positives
1	266	254	12	31
2	217	195	22	45
3	48	43	5	26

Table 1: Results on Real Datasets

(a) mailorder, $\geqslant 0.2\%$ true support

		predicted support			
size	Itemsets	< 0.1	0.1−0.15	0.15−0.2	$\geqslant 0.2$
1	65	0	0	0	65
2	228	0	1	15	212
3	22	0	1	3	18

(b) soccer, $\geqslant 0.2\%$ true support

		predicted support			
size	Itemsets	< 0.1	0.1−0.15	0.15−0.2	$\geqslant 0.2$
1	266	0	2	10	254
2	217	0	5	17	195
3	48	0	1	4	43

Table 2: Analysis of false drops

(a) mailorder, $\geqslant 0.2\%$ predicted support

		true support			
size	Itemsets	< 0.1	0.1−0.15	0.15−0.2	$\geqslant 0.2$
1	65	0	0	0	65
2	240	0	0	28	212
3	23	1	2	2	18

(b) soccer, $\geqslant 0.2\%$ predicted support

		true support			
size	Itemsets	< 0.1	0.1−0.15	0.15−0.2	$\geqslant 0.2$
1	285	0	7	24	254
2	240	7	10	28	195
3	69	5	13	8	43

Table 3: Analysis of false positives

[5] If there are no frequent itemsets for some combination, we pick the itemsets with the highest support.

[6] While we could have easily corrected the slip, we felt it more instructive to leave it in.

soccer

Transaction size:	1	2	3	4	5	6	7	8	9	10
Worst Average:	1: 4.4%	2: 20.2%	3: 39.2%	4: 44.5%	4: 43.9%	4: 37.5%	4: 36.2%	4: 38.7%	8: 51.0%	10: 49.4%
Worst of the Worst:	1: 45.5%	2: 45.4%	3: 53.2%	4: 49.8%	5: 49.7%	5: 42.7%	5: 41.8%	5: 44.5%	9: 66.2%	10: 65.6%

mailorder

Transaction size:	1	2	3	4	5	6	7
Worst Average:	1: 12.0%	2: 27.5%	3: 48.4%	4: 51.5%	5: 51.7%	5: 51.9%	6: 49.8%
Worst of the Worst:	1: 47.6%	2: 51.9%	3: 53.6%	4: 53.1%	5: 53.6%	6: 55.4%	7: 51.9%

Table 4: Actual Privacy Breaches

6. CONCLUSIONS

In this paper, we have presented three key contributions toward mining association rules while preserving privacy. First, we pointed out the problem of privacy breaches, presented their formal definitions and proposed a natural solution. Second, we gave a sound mathematical treatment for a class of randomization algorithms and derived formulae for support and variance prediction, and showed how to incorporate these formulae into mining algorithms. Finally, we presented experimental results that validated the algorithm in practice by applying it to two real datasets from different domains.

We conclude by raising three interesting questions for future research. Our approach deals with a restricted (albeit important) class of privacy breaches; can we extend it to cover other kinds of breaches? Second, what are the theoretical limits on discoverability for a given level of privacy (and vice versa)? Finally, can we combine randomization and cryptographic protocols to get the strengths of both without the weaknesses of either?

7. REFERENCES

[1] N. R. Adam and J. C. Wortman. Security-control methods for statistical databases. *ACM Computing Surveys*, 21(4):515–556, Dec. 1989.

[2] D. Agrawal and C. C. Aggarwal. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In *Proc. of the 20th ACM Symposium on Principles of Database Systems*, pages 247–255, Santa Barbara, California, May 2001.

[3] R. Agrawal. Data Mining: Crossing the Chasm. In *5th Int'l Conference on Knowledge Discovery in Databases and Data Mining*, San Diego, California, August 1999. Available from `http://www.almaden.ibm.com/cs/quest/papers/kdd99_chasm.ppt`.

[4] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 207–216, Washington, D.C., May 1993.

[5] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast Discovery of Association Rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, chapter 12, pages 307–328. AAAI/MIT Press, 1996.

[6] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1994.

[7] R. Agrawal and R. Srikant. Privacy preserving data mining. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 439–450, Dallas, Texas, May 2000.

[8] R. Bayardo. Efficiently mining long patterns from databases. In *Proc. of the ACM SIGMOD Conference on Management of Data*, Seattle, Washington, 1998.

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, 1984.

[10] Business Week. *Privacy on the Net*, March 2000.

[11] C. Clifton and D. Marks. Security and privacy implications of data mining. In *ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 15–19, May 1996.

[12] R. Conway and D. Strip. Selective partial access to a database. In *Proc. ACM Annual Conf.*, pages 85–89, 1976.

[13] L. Cranor, J. Reagle, and M. Ackerman. Beyond concern: Understanding net users' attitudes about online privacy. Technical Report TR 99.4.3, AT&T Labs–Research, April 1999.

[14] L. F. Cranor, editor. *Special Issue on Internet Privacy.* Comm. ACM, 42(2), Feb. 1999.

[15] The Economist. *The End of Privacy*, May 1999.

[16] V. Estivill-Castro and L. Brankovic. Data swapping: Balancing privacy against precision in mining for logic rules. In M. Mohania and A. Tjoa, editors, *Data Warehousing and Knowledge Discovery DaWaK-99*, pages 389–398. Springer-Verlag Lecture Notes in Computer Science 1676, 1999.

[17] European Union. *Directive on Privacy Protection*, October 1998.

[18] Y. Lindell and B. Pinkas. Privacy preserving data mining. In *CRYPTO*, pages 36–54, 2000.

[19] T. M. Mitchell. *Machine Learning*, chapter 6. McGraw-Hill, 1997.

[20] Office of the Information and Privacy Commissioner, Ontario. *Data Mining: Staking a Claim on Your Privacy*, January 1998.

[21] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[22] A. Shoshani. Statistical databases: Characteristics, problems and some solutions. In *VLDB*, pages 208–213, Mexico City, Mexico, September 1982.

[23] K. Thearling. Data mining and privacy: A conflict in making. *DS**, March 1998.

[24] Time. *The Death of Privacy*, August 1997.

[25] J. Vaidya and C. W. Clifton. Privacy preserving

association rule mining in vertically partitioned data. In *Proc. of the 8th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, July 2002.

[26] S. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.*, 60(309):63–69, March 1965.

[27] A. Westin. E-commerce and privacy: What net users want. Technical report, Louis Harris & Associates, June 1998.

[28] A. Westin. Privacy concerns & consumer choice. Technical report, Louis Harris & Associates, Dec. 1998.

[29] A. Westin. Freebies and privacy: What net users think. Technical report, Opinion Research Corporation, July 1999.

APPENDIX

A. PROOFS

A.1 Proof of Statement 1

Proof. Each coordinate $N \cdot s'_{l'}$ of the vector in (4) is, by definition of partial supports, just the number of transactions in the randomized sequence T' that have intersections with A of size l'. Each randomized transaction t' contributes to one and only one coordinate $N \cdot s'_{l'}$, namely to the one with $l' = \#(t' \cap A)$. Since we are dealing with a per-transaction randomization, different randomized transactions contribute independently to one of the coordinates. Moreover, by item-invariance assumption, the probability that a given randomized transaction contributes to the coordinate number l' depends only on the size of the original transaction t (which equals m) and the size l of intersection $t \cap A$. This probability equals $p\,[l \to l']$.

So, for all transactions in T that have intersections with A of the same size l (and there are $N \cdot s_l$ such transactions) the probabilities of contributing to various coordinates $N \cdot s'_{l'}$ are the same. We can split all N transactions into $k+1$ groups according to their intersection size with A. Each group contributes to the vector in (4) as a multinomial distribution with probabilities

$$(p\,[l \to 0]\,, p\,[l \to 1]\,, \ldots\,, p\,[l \to k]),$$

independently from the other groups. Therefore the vector in (4) is a sum of $k+1$ independent multinomials. Now it is easy to compute both expectation and covariance.

For a multinomial distribution $(X_0, X_1, \ldots, X_k)$ with probabilities $(p_0, p_1, \ldots, p_k)$, where $X_0 + X_1 + \ldots + X_k = n$, we have $\mathbf{E}\ X_i = n \cdot p_i$ and

$$\mathbf{Cov}\,(X_i, X_j) = \mathbf{E}\,(X_i - p_i)(X_j - p_j) = n \cdot (p_i \delta_{i=j} - p_i p_j).$$

In our case, $X_i = l$'s part of $N \cdot s'_i$, $n = N \cdot s_l$, and $p_i = p\,[l \to i]$. For a sum of independent multinomial distributions, their expectations and covariances add together:

$$\mathbf{E}\,(N \cdot s'_{l'}) = \sum_{l=0}^{k} N \cdot s_l \cdot p\,[l \to l']\,,$$

$$\mathbf{Cov}\,(N \cdot s'_i, N \cdot s'_j) =$$

$$= \sum_{l=0}^{k} N \cdot s_l \cdot (p\,[l \to i] \cdot \delta_{i=j} - p\,[l \to i] \cdot p\,[l \to j])$$

Thus, after dividing by an appropriate power of N, the formulae in the statement are proven. □

A.2 Proof of Statement 2

Proof. We are given a transaction $t \in T$ and an itemset $A \subseteq \mathcal{I}$, such that $|t| = m$, $|A| = k$, and $\#(t \cap A) = l$. In the beginning of randomization, a number j is selected with distribution $\{p_m[j]\}$, and this is what the first summation takes care of. Now assume that we retain exactly j items of t, and discard $m - j$ items.

Suppose there are q items from $t \cap A$ among the retained items. How likely is this? Well, there are $\binom{m}{j}$ possible ways to choose j items from transaction t; and there are $\binom{l}{q}\binom{m-l}{j-q}$ possible ways to choose q items from $t \cap A$ and $j - q$ items from $t \setminus A$. Since all choices are equiprobable, we get $\binom{l}{q}\binom{m-l}{j-q}/\binom{m}{j}$ as the probability that exactly q A-items are retained.

To make t' contain exactly l' items from A, we have to get additional $l' - q$ items from $A \setminus t$. We know that $\#(A \setminus t) = k - l$, and that any such item has probability ρ to get into t'. The last terms in (8) immediately follow. Summation bounds restrict q to its actually possible (= nonzero probability) values. □

A.3 Proof of Statement 3

Proof. Let us denote

$$\vec{p_l} := (p\,[l \to 0]\,, p\,[l \to 1]\,, \ldots\,, p\,[l \to k])^T,$$
$$\vec{q_l} := (q\,[l \leftarrow 0]\,, q\,[l \leftarrow 1]\,, \ldots\,, q\,[l \leftarrow k])^T.$$

Since $PQ = QP = I$ (where I is the identity matrix), we have

$$\sum_{l=0}^{k} p\,[l \to i]\ q\,[l \leftarrow j] = \sum_{l'=0}^{k} p\,[i \to l']\ q\,[j \leftarrow l'] = \delta_{i=j}.$$

Notice also, from (7), that matrix $D[l]$ can be written as

$$D[l] = \operatorname{diag}(\vec{p_l}) - \vec{p_l}\,\vec{p_l}^{\,T},$$

where $\operatorname{diag}(\vec{p_l})$ denotes the diagonal matrix with $\vec{p_l}$-coordinates as its diagonal elements. Now it is easy to see that

$$\tilde{s} = \vec{q}_k^{\;T} \vec{s}\,' = \sum_{l'=0}^{k} q\left[k \leftarrow l'\right] \cdot s'_{l'};$$

$$\begin{aligned}
\mathbf{Var}\;\tilde{s} &= \frac{1}{N} \sum_{l=0}^{k} s_l\, \vec{q}_k^{\;T} D[l]\, \vec{q}_k = \\
&= \frac{1}{N} \sum_{l=0}^{k} s_l\, \vec{q}_k^{\;T} \left(\mathrm{diag}(\vec{p}_l) - \vec{p}_l\, \vec{p}_l^{\;T}\right) \vec{q}_k = \\
&= \frac{1}{N} \sum_{l=0}^{k} s_l \left(\vec{q}_k^{\;T} \mathrm{diag}(\vec{p}_l)\, \vec{q}_k - (\vec{p}_l^{\;T} \vec{q}_k)^2\right) = \\
&= \frac{1}{N} \sum_{l=0}^{k} s_l \Big(\sum_{l'=0}^{k} p\left[l \rightarrow l'\right]\, q\left[k \leftarrow l'\right]^2 - \delta_{l=k}\Big);
\end{aligned}$$

$$\begin{aligned}
&(\mathbf{Var}\;\tilde{s})_{\mathrm{est}} \;= \\
&= \frac{1}{N} \sum_{l=0}^{k} (\vec{q}_l^{\;T} \vec{s}\,') \Big(\sum_{l'=0}^{k} p\left[l \rightarrow l'\right]\, q\left[k \leftarrow l'\right]^2 - \delta_{l=k}\Big) = \\
&\quad = \frac{1}{N} \sum_{j=0}^{k} s'_j \Big(\sum_{l,l'=0}^{k} q\left[l \leftarrow j\right] p\left[l \rightarrow l'\right]\, q\left[k \leftarrow l'\right]^2 - \\
&- \sum_{l=0}^{k} \delta_{l=k}\, q\left[l \leftarrow j\right]\Big) \;=\; \frac{1}{N} \sum_{j=0}^{k} s'_j \Big(\sum_{l'=0}^{k} \delta_{l'=j}\, q\left[k \leftarrow l'\right]^2 - \\
&\qquad - q\left[k \leftarrow j\right]\Big) \;=\; \frac{1}{N} \sum_{j=0}^{k} s'_j \big(q\left[k \leftarrow j\right]^2 - q\left[k \leftarrow j\right]\big).
\end{aligned}$$

□

A.4 Proof of Statement 4

Proof. We prove the left formula in (13) first, and then show that the right one follows from the left one. Consider $N \cdot \Sigma_l$; it equals

$$\begin{aligned}
N \cdot \Sigma_l = N \cdot \sum_{C \subseteq A,\ |C| = l} \mathrm{supp}^T(C) &= \sum_{C \subseteq A,\ |C| = l} \#\{t_i \in T \mid C \subseteq t_i\} = \\
&= \sum_{i=1}^{N} \#\{C \subseteq A \mid |C| = l, C \subseteq t_i\}.
\end{aligned}$$

In other words, each transaction t_i should be counted as many times as many different l-sized subsets $C \subseteq A$ it contains. From simple combinatorics we know that if $j = \#(A \cap t_i)$ and $j \geqslant l$, then t_i contains $\binom{j}{l}$ different l-sized subsets of A. Therefore,

$$\begin{aligned}
&N \cdot \Sigma_l = \sum_{i=1}^{N} \binom{\#(A \cap t_i)}{l} = \\
&= \sum_{j=l}^{k} \binom{j}{l} \cdot \#\{t_i \in T \mid \#(A \cap t_i) = j\} = \sum_{j=l}^{k} \binom{j}{l} N \cdot s_j,
\end{aligned}$$

and the left formula is proven. Now we can check the right formula just by replacing the Σ_j's according to the left formula. We have:

$$\begin{aligned}
&\sum_{j=l}^{k} (-1)^{j-l} \binom{j}{l} \Sigma_j = \sum_{j=l}^{k} (-1)^{j-l} \binom{j}{l} \sum_{q=j}^{k} \binom{q}{j} s_q = \\
&= \sum_{l \leqslant j \leqslant q \leqslant k} (-1)^{j-l} \binom{j}{l} \binom{q}{j} s_q = \sum_{q=l}^{k} s_q \sum_{j=l}^{q} (-1)^{j-l} \binom{j}{l} \binom{q}{j} \\
&\quad = \sum_{q=l}^{k} s_q \sum_{j'=0}^{q-l} (-1)^{j'} \frac{(j'+l)!}{l!\, j'!}\, \frac{q!}{(j'+l)!\,(q-j'-l)!} = \\
&\qquad = \sum_{q=l}^{k} s_q \cdot \frac{q!}{l!\,(q-l)!} \sum_{j'=0}^{q-l} (-1)^{j'} \frac{(q-l)!}{j'!\,(q-l-j')!} = \\
&\qquad\qquad = \sum_{q=l}^{k} s_q \binom{q}{l} \sum_{j'=0}^{q-l} (-1)^{j'} \binom{q-l}{j'} = s_l,
\end{aligned}$$

since the sum $\sum_{j'=0}^{q-l} (-1)^{j'} \binom{q-l}{j'}$ is zero whenever $q-l > 0$.

To prove that matrix P becomes lower triangular after the transformation from $\vec{s}$ and $\vec{s}\,'$ to $\vec{\Sigma}$ and $\vec{\Sigma}\,'$, let us find how $\mathbf{E}\;\vec{\Sigma}\,'$ depends on $\vec{\Sigma}$ using the definition (12).

$$\begin{aligned}
&\mathbf{E}\;\Sigma'_{l'} = \sum_{C \subseteq A,\ |C| = l'} \mathbf{E}\;\mathrm{supp}^{T'}(C) = \\
&\quad = \sum_{C \subseteq A,\ |C| = l'} \sum_{l=0}^{l'} p_{l'}^{m}\left[l \rightarrow l'\right] \cdot \mathrm{supp}_l^{T}(C) = \\
&= \sum_{C \subseteq A,\ |C| = l'} \sum_{l=0}^{l'} p_{l'}^{m}\left[l \rightarrow l'\right] \sum_{j=l}^{l'} (-1)^{j-l} \binom{j}{l} \Sigma_j(C,T) = \\
&= \sum_{j=0}^{l'} \underbrace{\sum_{l=0}^{j} (-1)^{j-l} \binom{j}{l} p_{l'}^{m}\left[l \rightarrow l'\right]}_{c_{l'\,j}} \sum_{C \subseteq A,\ |C| = l'} \Sigma_j(C,T) = \\
&\quad = \sum_{j=0}^{l'} c_{l'\,j} \sum_{C \subseteq A,\ |C| = l'} \;\sum_{B \subseteq C,\ |B| = j} \mathrm{supp}^T(B) = \\
&= \sum_{j=0}^{l'} c_{l'\,j} \sum_{B \subseteq A,\ |B| = j} \#\{C \mid B \subseteq C \subseteq A, |C| = l'\} \cdot \mathrm{supp}^T(B) = \\
&= \sum_{j=0}^{l'} c_{l'\,j} \sum_{B \subseteq A,\ |B| = j} \binom{k-j}{l'-j} \mathrm{supp}^T(B) = \sum_{j=0}^{l'} c_{l'\,j} \binom{k-j}{l'-j} \cdot \Sigma_j.
\end{aligned}$$

Now it is clear that only the lower triangle of the matrix can have non-zeros. □