An Appraisal on Privacy Preserving Mining of Association Rules

C. Anitha
Assistant Professor,
Department of MCA,
C.R.Engineering College,
Tirupati, India

Prof. M. Padmavathamma
Head, Department of Computer
Science, S.V.University,
Tirupathi, India

M. Sunil Kumar
Associate Professor
C R Engineering College,
Tirupati, India

ABSTRACT

An interesting new direction for data mining research is the development of techniques that incorporate privacy concerns for association rules. In this work, we present a framework for mining association rules from various transactions. These transactions mainly consisting of categorical items, where the data has to preserve privacy of individual transactions. By using uniform randomization, it is feasible to recover association rules, but these rules are in turn be exploited to find privacy breaches. Hence, in this work we clearly analyze the nature of privacy breaches and propose a new class of randomization operators that are much more effective than uniform randomization which was proposed previously. Here we also derive formulae for an unbiased support estimator, which allows us to recover item set supports from randomization data sets. Here we also show how the above derived formulae will be incorporated into mining algorithms. Finally; we provide experimental results that validate the proposed algorithm by applying it to real data sets.

1. INTRODUCTION

It is estimated tht amount of information in world is doubling for every 20 months since there is an explosive progress in networking, storage & process technologies. This result in an unpredicted amount of digitization of information. With the dramatic increase in digital data, concerns about privacy of personal information have emerged globally [15][17][20][24]. Privacy issues are once again increasing because internet makes it easy for the new data to be automatically collected and added to databases[10][13][14]. By using data mining we will efficiently discover valuable, non-obvious information from large databases. Hence, we will make obvious use to data mining in an interesting new direction which is vulnerable to misuse[11][16].

The new direction in datamining research is the development of techniques that incorporates privacy breaches [3]. The following point to be remembers when using datamining in new direction: The primary task in datamining is development of models about aggregated data, but it is not possible to 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 ina users record have been randomized so that true values cannot be estimated with sufficient precision. Here randomization is done using the statistical method of value distortion that returns a value xi+r instead of xi where r is random value drawn from some distribution. For correcting perturbed distributing, they proposed a Bayesian procedure and presented 3 algorithms for building accurate decision trees [9][21] which mainly relys on reconstructed distributions. In [2], authors derived Expectation maximization (EM) algorithm for reconstruction distributions and prove that EM algorithm converged to the maximum likelihood estimate of the original distribution based on perturbed data. They also prove that EM algorithm was in fact identical to the Bayesian reconstruction procedure in [7], expect for an approximation that made that at later time.

1.1 Contribution

We will use the reconstruction in developing privacy-preserving data mining techniques and extend this inquiry mainly in two dimensions:

- Categorical data and
- Association rule mining

We will mainly focus on the task of finding frequent itemsets in association rule mining;

Suppose we have a set of I of n items:I={a1,a2,a3.......an}. Let T be a sequence of N Transaction T=(t1,t2,....tn) where each transaction ti is a subset of I. Given an item set ACI, its support supp^T (A) is defined as

$$\text{supp}^T(A)= \frac{\#\{t\in T\}\ a\leq t\}}{N} \quad (1)$$

An item set ACI is called frequent in T if

$\text{supp}^T(A)\geq T$, where T is user-defined parameter.

Consider a scenario where we have a server and many clients. Each client has set of items where a client wants server together statistical information about associations among items, in order to provide recommendations to the clients. Normally when server get set of items from clients, it modifies according to some randomization policy. The server then gathers statistical information from the modified set of items and recovers from it the actual associations.

The remaining work in the paper proceeds as follows: In section 2, we show the uniform randomization which leads to privacy breaches. We define privacy breaches in section 3. In the section 4 we discuss about various randomization operators that can be tuned for different tradeoffs between discoverability & privacy breaches. In section 5, we show the experimental results on two real data sets, as well as graphs showing the relationship between discoverability, privacy & data characteristics.

2. RELATED WORK

By the desire to provide statistical information extensive research has been done in the area of statistical information without comprising sensitive information about individuals. The proposed technique has been broadly classified in to query restricted & data perturbation. Query restriction family includes restricting the size of query result, controlling overlap amongst successive queries, keeping a list of all answered queries and checking for possible compromise, suppression of data cells of small size and clustering entites 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 result of a query. Negative results showing in the proposed technique cannot satisfy the conflicting objectives of providing high quality statistics and at the same time prevent exact disclosure of individual information[1].

The more relevant work for the statistical data was done by Warner where he developed the "randomized response" the method for survey results. The whole approach will be dealed with a single Boolean attribute. This approach may be viewed as a general Warner's idea. Another related work is, where they consider the problem of mining association rules over data i.e vertically partitioned across two sources i.e, for each transaction, some of the item sets based on multi-party computation technique for scalar products, without either sources revealing exactly which transaction support a subset of item sets. We mainly focus on preserving privacy when the data is horizontally partitioned, we can preserve privacy for individual transaction rather than between two data sources.

Another related work [16],is the problem of inducing decision tress over horizontally partitioned training data which may not trust others. Here each source first builds a local decision tree over its true data, and then swaps the values over the leaf node of the tree to generate randomized training data. Another approach [18], do not use randomization, but makes use of cryptographic obvious functions during tree construction of data sources.

3. UNIFORM RANDOMIZATION

Normal approach for randomizing transaction would be to generalize "warners" randomized response" method which is described in the previous section. When a server sends a transaction to the server, the client takes each item and with probability P replaces it by a new item not originally present in the transaction. This randomization call is called as uniform randomization.

Estimating an item set is nontrivial even for uniform randomization. Randomized support depends not only on its true support, but also on the support of subsets. In these 3-itemsets, one or two of the items are inserted by chance than all three. Hence, almost all 'false' occurrences of the item sets are due to high subset supports. This requires estimating the support of all subsets simultaneously for large values of P; most of the items in randomized transaction may be "false", so that we obtain a reasonable privacy protection. Also , if there are enough clients and transaction, then frequent item sets will still be "visible", though similarly frequency than originally. This randomization has problem. If we know that our 3-item set escapes randomization in so per million transaction, and that even occurs because of randomization, then every time we it in a randomized transaction, with even more certainly at least one itemset from this item sets will be true i.e a chance insertion of one or two the items is much more likely than of all three. In this case we 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 of high values of P. The remaining paper address design a frame work for studying privacy breaches and developing techniques for finding frequent item sets while avoiding breaches.

3.1 Privacy breaches:-

Definition 2: let (Ω, f, P) be a probability space of elementary events over some set Ω and σ- algebra f. A randomization operator is a measurable function

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

That randomly transforms a sequence of N transaction into a different sequence of N transactions. In a given sequence of n transactions T, we shall write $T^1 = R(T)$ where T is a constant and R(T) is a random variable.

Definition 3:-

A general privacy breach of level p with respect to a property $p(t_i)$ occurs if

$$\exists \quad T^1 : P[P(t_i) \mid R(T)=T^1] \geq P$$

We say that a property $Q(T^1)$ causes a privacy breach of level p with reference to $P(t_i)$ if

$$P[P(T_i) \mid Q(R[T]))] \geq P$$

If we have know about a prior distribution, then we define privacy breach, so that it makes sense to speak about a posterior probability of property $P(t_i)$ versus prior. In prior distribution, transactions are not randomly generated. When we have modeling transactions as being randomly generated from a prior distribution which allows us to clearly define a privacy breach?

Definition 4:-

We say that item set A causes a privacy breach of level P if for some item $a \in A$ and some $i \in 1 \ldots N$ we have $P[\alpha \in t_i \mid A \leq t_i^{\ 1}] \geq P$

Here we ignore the effect of other information that server obtains from randomized transactions in which items the randomized transactions does not contain the information about the breaches has been known to the server in prior and also it knows other information about items and clients besides the transactions. In some scenarios, being confident that an item was not present in the original transaction may also be considered as a privacy breach.

4. ALGORITHM

Definiton5:-

We call randomization R a per-transaction randomization if, for $T=(t_1,t_2,t_3,\ldots.t_N)$ we can represent R(T) as
$R(t_1,t_2,t_3,\ldots.t_N)=(R(1, t_1),R(2, t_2),\ldots.R(N,t_N))$,
Where R(i,t) are independent random variable whose distributions depends only on t. we shall write $t_i^1=R(i, t_1)=R(t_1)$.

Definition 6:-

A randomization operator R is called invariant if, for every transaction sequence T and for every permutation π: I→ I of items, the distribution of
$\Pi^{-1}\ R(\pi\ ^{T})$ is the same as of R(T). here $\pi\ ^{T}$ 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 operators, for each possible input transaction size m:
Default probability of an item pm Є (0,1);
Sum of transaction subset size selection probabilities $P_m[0],P_m[1]\ldots.P_m[m]$ such that $P_m[T]\geq 0$, is equal to 1.
$P_m[0]+P_m[1]+\ldots.+P_m[m]=1$

Given a sequence of transaction $T=(t_1,t_2,\ldots.t_N)$, the operator takes each transaction t_i independently and proceeds as follows to obtain t_i^1.

(1) The operator selects an integer J at random from the set {0,1….,m} so that
P[J is selected]=$P_m[j]$

(2) It select j items from t_i, uniformly at random, hence no other items of t_i, are placed into t_i^1.

(3) It consider each item a ∉ t_i in turn and all those items are added to t_i^1.

Definition 8:-

A cut-and-paste randomization operator is a special case of a select-a-size operator. For each possible input transaction size m, it has two parameters: P_m Є (0,1) and an intent $k_m>0$. The operator takes each input transaction $Є_i$ independently and precedes as follows to obtain transaction $t_i^{1.}$

(1) It chooses an integer J uniformly at random between o and K_m; if j>m, it sets j=m.

(2) The operator select items out of t_i uniformly at random. These items are placed into t_i^1.

(3) Each other item is placed into t_i^1cwith probability P_m, independently.

The mixing randomization operator has one integer parameter k>=2 and one real-valued parameter pЄ(0,10. If we are having transactions $T=(\ t_1,t_2,t_3,\ldots.t_N)$ the operator takes each transaction t_i independently and proceeds as follows to obtain t_i^1.

(1) Other than t_i, pick K⇔1 more transaction from T and union the K transaction as sets of items. Let t_i^{11}be the union.

(2) Consider each item as a aЄ t_i^{11} with probability P.

(3) All those items other than the required probability are removed.

Privacy preserving data-mining, mostly focuses on per-transaction randomizations, since they are easiest & safest to implement. Here users does not communicate with each other, nor they cannot exchange random bits. Hence implementing mixes randomization, requires to organize an exchange of non-randomized transactions between users.

4.2 Effects of Randomization:-

In randomization, let T be a sequence of transactions of length N, and let A be some subset of items. Suppose we randomize T and get $T^1=R(T)$. hence, suppose $S^1=$ Supp $^{T1}(A)$ of A for T^1 is a random variable that depends on the outcome of the randomizations. 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 transaction in T that have intersection with A of size l among all transaction in T is called partial support of A for intersection of size l;

$$\mathrm{Supp}_l^{\,T}=\frac{\#\{t\in T\,|\,\#(A\cap t)=l}{N} \qquad (2)$$

It is easy to see that Supp $^T(A)=\mathrm{Supp}^T{}_K(A)$ for k=|A|, and that

$$\sum_{l=0}^{K}\ \mathrm{Supp}_l^{\,T}(A)=1$$

Since those transaction in T that do not intersect A at all are covered in $\mathrm{Supp}_0^{\,T}(A)$

Definition 10:-

Suppose that our randomization operator is both per-transaction and item-invariant. Consider a transaction t of size m and an item set A c I of size K. After randomization, transaction t becomes t. we define

$$P_k\,^m[l\rightarrow i]=P[l\rightarrow i]:=P[\#(T^1\cap A)=i|\#(t\cap A)=l]\quad (3)$$

Here both l and I must be integers in {0,1,….k}

STATEMENT 1:-

Suppose that our randomization operator ids both per-transaction and item-invariant all the N transaction in T have the same size m. then for given subset equation

$$N(S_0^1,S_1^1,S_2^1,\ldots\ldots.S_K^1);\text{ where } S_1^1=\mathrm{supp}_1^{\,T1}(A)\ (4)$$

Is a sum of K+1 independent random vector, each having a multinomial distribution. Its expected value is given by

$$E(S_0^1,S_1^1,S_2^1,\ldots\ldots.S_K^1)=p(S_0,S_1,\ldots.S_k)^T \qquad (5)$$

Where p is the (K+1)x(k+1) matrix with elements Pi,l=P[l→ i], and the covariance matrix is given by

$$\mathrm{Cov}(S_0^1,S_1^1,S_2^1,\ldots\ldots.S_K^1)=\frac{1}{N}\cdot\sum_{i=0}^{k}\ S_l\,D[l] \quad (6)$$

Where each D[l] is a (k+l)X(k+l) matrix with elements

$D[l]_{ij}=p[l\rightarrow i].S_{i=j} \Leftrightarrow P[l\rightarrow i].P[l\rightarrow j]$ (7)

Here S_l denotes $\mathrm{supp}_l^T(A)$, and the T over vector denotes the transpose operation; $S_{i=j}$ is one if i=j and zero otherwise.

Support Recovery:

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

$\overrightarrow{s}$: $(S_{0,}S_{1,\ldots.}S_k)^T$, $S^1=(S_{0,}S_{1,\ldots.}S_k)^T$; (8)

According to 5 we have E. $\overrightarrow{s}$= p. $\overrightarrow{s}$ (9)

Denote $Q=P^{-1}$ and multiply both sides by (9) by Q:

$\overrightarrow{s} = Q.E\overrightarrow{s} = E.Q.\overrightarrow{s}$

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

$\overrightarrow{s}_{est}=Q.\overrightarrow{s}$ (10)

Using (6), we can compute the covariance matrix of $\overrightarrow{s}_{est.}$

$\mathrm{Cov}\,\overrightarrow{s}_{est}=\mathrm{cov}(Q.\overrightarrow{s})=Q(\mathrm{Cov}.\overrightarrow{s}) = \frac{1}{N}.\sum_{l=0}^{K} S_l\, QD[l]Q^T$

(11)

This estimator is also unbiased:

$E(\mathrm{Cov}\,\overrightarrow{s}_{est})_{est}=\frac{1}{N}.\sum_{l=0}^{K} (E\overrightarrow{s}_{est})_l\, Q\, D[l]Q^T= \mathrm{Cov}.\,\overrightarrow{s}_{est.}$

Definition 11:-

Suppose we have a transaction sequence T and an item set **A I.** Given an integer L between 0 and K= |A|, consider all subsets **C A** of size l. The sum of support of all these subsets is called the cumulative support for A of order l and is denoted as follows:

$\sum_l = \sum_l (A,T):=\sum_{C\leq A,|c|=l} \mathrm{Supp}^T(C)$

, $\overrightarrow{\Sigma}=(\Sigma_{0,}\ \Sigma_{1,}\ \ldots..\Sigma_{K})^T$ (12)

Limiting privacy Breaches:

In this section we determine how privacy depends on randomization we use definitions and assume a per-transaction and item-invariant randomization.

Consider some itemset $A\leq I$ and some item $a\in A$, f in 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 non randomized sizes. Assume also that a particular support $S_l=\mathrm{Supp}_l^T(A)$ approximate the corresponding prior probabilities $p[\#(t\cap A)=l]$. suppose we know the following prior probabilities:

$S_l^+=p[\#(t\cap A)=l,a\in t]$.

$S_l^-=p[\#(t\cap A)=l,a\notin t]$.

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

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

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

$\sum_{i=0}^{K} S_l^+. P[l\rightarrow k]<0.5 \sum_{i=0}^{K} S_{l}.P[l\rightarrow k]$ (14)

The problem is that we have to randomize the data before we know any supports. Also we may not have the luxary of selting “opver safe” randomization parameter because then we may not here a 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 $_{K-I}$ itemset in the transaction of given size m, for different k& m.

(2) Given the maximum support, find values for S_l and S_l^+ that are more likely to cause a privacy breach.

4.3 Discovering Associations:

Given a set of randomized transaction, we show how to discover itemset with high true support. We mainly use the Apriori algorithm [1] to make the ideas concentrate, all the modifications will be applied directly to any algorithm that use Apriori candidate generation. The key lattice property of our Apriori is that, for any two itemset $A\leq b$, the true support of A’s equal to (or) larger than the true support of B. Simplest version of Apriori algorithm 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) Compute all candidate sets whose supports low S_{min}.
  (c) Save the remaining candidate set for output;
  (d) Form all possible (k+1)-item sets. Such that all their k-subsets are among the remaining candidates. Let these itemset are new candidate set.

(2) Output are saved item sets.

We can directly modify this algorithm so that it reads the randomized dataset, compute partial support of all candidate and recovers sigmas using formulae from statements 3. Normally we will fix a minimum support for all data sets, if we discard all candidate, below minimum support for the purpose of candidate generation, we will miss many of the longer frequent item sets. Here I the modified version of the Apriori:

(1) Let k=1, let “ candidate set” be all single-item sets, Repeat the following until K is too large for support recovery.
  (a) Read the randomized data file and compute the partial supports of all candidate sets, separately for each non randomized transaction size;
  (b) Recover the predicted supports and sigmas for the candidate sets.
  (c) Discard every candidate set whose support is low to tis candidate limit.
  (d) Save for uoutput only those candidate set whose predicated support is atleast S_{min};
  (e) From all possible (k+1)- item sets such that all their k-subsets are among the remaining candidates. Let all the itemset be new candidate sets.

(f) Let K=K+1

(2) Output all the saved itemsets.

5. EXPERIMENTAL RESULTS

In this section we show our ability to recover supports depends on the permitted breach level. Next we will describe real-time dataset and present result on these datasets in last section .

5.1 Dataset characteristics:

Here we define the lowest discoverable support as the support at which the predicted support of an item set is four sigmas away from zero. In practice, we may achieve reasonably good results even if the minimum support level is slightly lower than four sigma. The lowest discoverable support is a nice way to illustrate the interaction between discoverability, privacy breach loss & data characteristics.

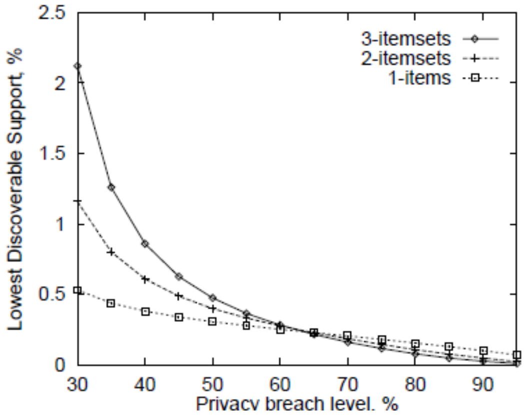


Figure 1 : Lowest discoverable support for different breach levels. Transaction size is 5, million transaction.

Figure 1 shows how the lowest discoverable support changes with privacy breach level. For higher privacy breach level at 95%, we can discover 3-itemsets at very low support. Similarly for lower privacy breach level at 50%, we discover itemset at very high support. But at higher breach level it get harder to discover 1-itemset support than 3-itemset supports. When we add fewer false items at higher breach levels, we generate so much fewer false 3-itemset positives than false 1-itemset positives than 3-itemsets get an advantage over single items.

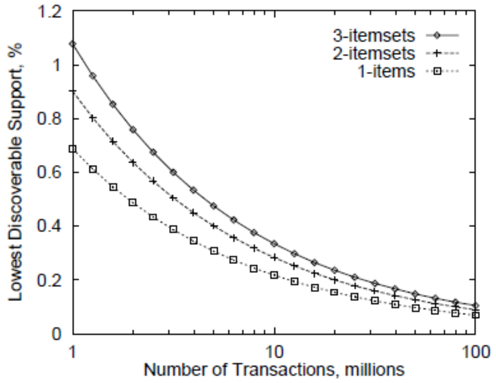


Figure 2: Lowest discoverable support versus number of transactions. Transaction size is 5, breach level is 50%.

Figure 2 shows lowest discoverable support is inversely proportional to the square root of the no of transactions. If all the partial support are fixed, the predictions variance is inversely proportional to the number N of transactions according to the statement 3.

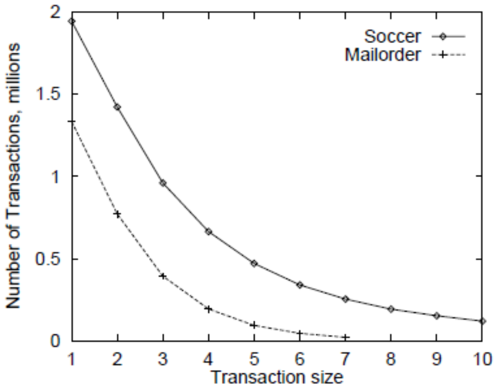


Figure 3 : Lowest discoverable support for different transaction sizes. Five million transactions, breach level is 50%

Figure 3 shows the transaction size has a significant 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 contain too much personal information to hide, because it may contain too much long itemset in the randomized transaction could result in a privacy breach. In such a long randomized transaction we have to insert a lot of false items and cut off many true ones to ensure that such a long items sets in randomized transaction. Such a strong randomization causes an exceedingly high variance in the support predictor for 2 and especially 3-itemsets.

Each item in the transaction is a web request. Not all the web request, were turned into item sets; to become an item, the request must satisfy the following

(1) client request method's GET

(2) Request status is ok

(3) File type is HTML.

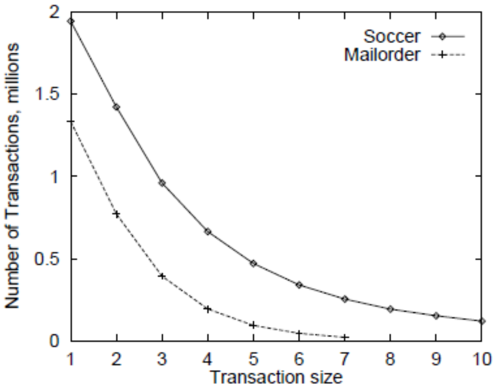


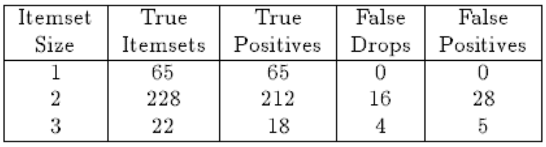
Figure 4: Number of transaction for each transaction size in the soccer and mail order datasets.

A session start with a request that satisfies the above properties, and ends when the last click from this client ID timeouts. All request in the session has same client ID. The soccer transaction file will proceeds further. After undergoing various processing, the resulting datasets, will be distributed as shown in the figure 4.

The Result:

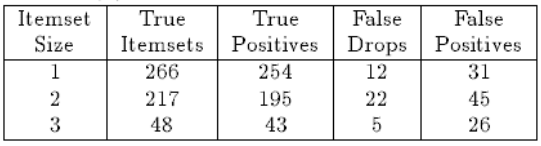
We show the result for datasets at a minimum support i.r close to the lowest discoverable support to show resilience of our algorithm at very low support levels. We mainly use cut-and-paste randomization(see definition 8) which has only 2 parameters, randomization level & cut off, per each transaction size. We use a cut off value of 7. When we all given with values of maximum support, we use 4.4 methodology to find the low lowest randomization level such that the breach probability is still below the desired breach level.

(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

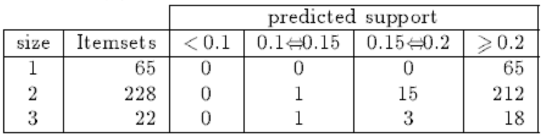
(b) Soccer 0.2% minimum support



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

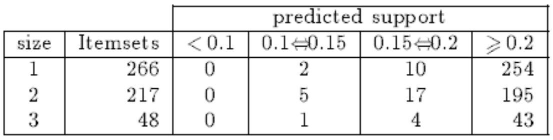
Table 1: Results on Real Datasets

(a) Mailorder, ≥ 0.2 % true support



		predicted support			
size	Itemsets	< 0.1	0.1⇔0.15	0.15⇔0.2	⩾ 0.2
1	65	0	0	0	65
2	228	0	1	15	212
3	22	0	1	3	18

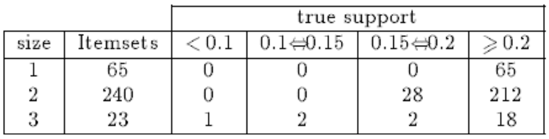
(b) Soccer, ≥ 0.2% true support



		predicted support			
size	Itemsets	< 0.1	0.1⇔0.15	0.15⇔0.2	⩾ 0.2
1	266	0	2	10	254
2	217	0	5	17	195
3	48	0	1	4	43

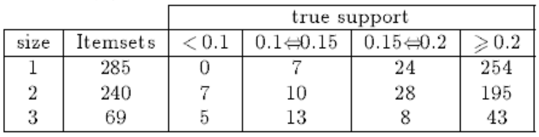
Table 2: Analysis of false drops

(a) Mailorder, ≥ 0.2 % predicted support



		true support			
size	Itemsets	< 0.1	0.1⇔0.15	0.15⇔0.2	⩾ 0.2
1	65	0	0	0	65
2	240	0	0	28	212
3	23	1	2	2	18

(b) Soccer, ≥ 0.2 % predicted support



		true support			
size	Itemsets	< 0.1	0.1⇔0.15	0.15⇔0.2	⩾ 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.

Table1 shows what happens fro the itemsets from both randomized & non randomized files and then compare the result. For lowest value also at 0.2 % most of the itemsets are mined correctly from the randomized file. There are comparatively fe false positives and even fewer false drops. The predicted sigma for 3-item set range from 0.066⇔ 0.07% and soccer in 0.047⇔ 0.048% for mail order. Since we know that there are more low-supported itemsets than there are highly supported, we might wonder most of the false ‘+’vs are outliers i.e, hare true support near zero. The table 2 & 3 show that usually the true supports of false ’+’vs , as well as the predicted support for false drops, are closer to 0.2% than to zero. This demonstrates randomization as a privacy-perserving approach.

6. CONCLUSION

In this paper, we proposed 3 key contributions for mining association rules. We can perform this contribution while preserving privacy. The first contribution specified the problem of privacy breaches, presented different formal definitions and proposed a best general solution. The second contribution provided mathematical treatment for different classes of randomized algorithm and derived formulae for variance prediction. Here we also shoed how to apply these formulae in mining algorithms. At finally we presented experimental result that validated the algorithm in practice from different domain.

We can conclude by raising some interesting questions for further research. Out paper deals with a single clay of privacy breachesican. We extend it to cover other kinds of breaches? Can we combine randomization & cryptographic protocols?

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 Quanti_cation 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] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classi_cation and Regression Trees. Wadsworth, Belmont, 1984.

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

[5] 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. [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.

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

[7] The Economist. The End of Privacy, May 1999.

[8] 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.

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

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

[11] J. R. Quinlan. Induction of decision trees. Machine Learning, 1:81{106, 1986.

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

[13] 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.

8. AUTHORS PROFILE

Mrs.C. Anitha has done her M.Sc., MCA and M.Phil, and pursuing her PhD in S.V.University, Tirupati. She has been an Assistant Professor in the Department of MCA, C.R.Engineering College, Tirupati teaching MCA, M.Sc and B.Tech students. Her areas of specialization are cryptography and network security, privacy preserving data mining.

Prof. M. Padmavathamma, M.Sc, M.S, M.Phil, M.Ed, Ph.D has been working as Head, Dept of Computer Science, in Sri Venkateswara University, Tirupati. She has 20 years of teaching experience for PG and 5 years for UG and has guided many PhD s. Her areas of specialization are cryptography & network security, privacy preserving data mining.