



Research Paper on Randomization-based Privacy-Preserving Association Rule Mining

Maiwand Khishki, Vijay Kumar
CSE Department, Chandigarh University
Chandigarh, India

Abstract— *this paper includes variable privacy levels in uniform randomization-based methods by varying the randomization coefficient. Another method is proposed that performs randomization at the user side, without storing the original data. This prevents the need for a trusted third party server this paper focuses only on methods where reconstruction of original data is not required. Finally, these methods are applied on a real dataset and experimental results are provided one of the many challenges of data mining is privacy and focuses on randomization methods for privacy preservation of categorical data. The aim of privacy preservation methods is to compute accurate results, without compromising the identity of the data owner. Different levels of privacy are an active research area in data mining. Different users may have Different needs for privacy*

Keywords— *Randomization, Association Rule Mining and privacy-preserving*

I. INTRODUCTION

1.1 PRIVACY PRESERVING DATA MINING

The rate at which data is generated is constantly increasing. With the availability of such large data and high computing power, data mining is being used extensively in many fields. One of the challenges of data mining is privacy [8]. Before data mining algorithms are run, data is sometimes intentionally modified from its original version. This is done to counter privacy and security threats. This distortion method is called randomization. This process of modification can distort the data to an unknown level. The accuracy of results obtained from distorted data is not known. The objective is to develop methods to find patterns from distorted data, and also to develop metrics to evaluate the integrity of these discovered patterns, while at the same time not to compensate the identity of the source(s)/owner(s) of the data. For privacy-preserving data mining, many methods are proposed in literature. Some of these methods are explained in the following section. This focuses on randomization based methods for mining categorical data. Randomization was first proposed in [3] on numerical data. For a data distribution X . Value distortion is performed on data record x_i by adding to it a random value r from a predefined random distribution R . All such values of $x_i + r$ are used to recreate the original distribution of Data using a Bayesian reconstruction procedure to perform data mining operations mining operations do not always need individual data. Having the distribution of data will be sufficient to run mining operations [3]. Let the randomized distribution $X + R$ is Y . The assumption is that it is easier to reconstruct the distribution X from Y . But difficult to reconstruct the individual values x_i (where; $i = 0; 1; 2; \dots$) from Y One factor to consider during randomization is the trade-off between accuracy and privacy [8]. The higher the need for accuracy, the lower is the privacy guarantee and vice-versa. Privacy is defined as the difficulty to identify individual rows of data which contributed to a given result. Certain attacks have been identified that could directly or indirectly lead to identifying the owner of the data. Though mining systems remove unwanted private data that could lead to identification of the data owner, there are attacks such as spectral analysis attack and public data attack by which privacy can be compromised [1]. Data collected to perform mining tasks are not always completely accurate. Especially when the data required is sensitive for the user. For example, in medical research, surveys that need data about past illness of people would have inaccurate data. People are apprehensive if such data would be used against them. Say a number of hospitals wish to predict the chances of a person contracting cancer based on their ailment history. Even if the individual's identity is not part of the survey, it is possible to identify their identities. Say the research group can collect billing information from medical shops that contain person name, time, and medicines bought. Or they can get data about doctors and a list of patients registered with them. By correlating such datasets, it could be possible to identify the individuals from whom data was collected.

1.2 CHALLENGES IN RANDOMIZATION-BASED METHODS

There are several challenges in randomization-based methods. Two significant challenges that are addressed in this paper are:

- Develop algorithms for comparing the two (original and randomized) versions of the data. This can be considered as privacy metric.
- Develop algorithms for estimating the impact that certain modifications of the data have on the statistical significance of individual patterns obtained by data mining algorithms.

The first challenge is to measure the amount of distortion that has been done and the corresponding privacy that has been introduced. And the second challenge is to measure the impact the modification of data values has on a discovered pattern's Accuracy. Creating a universal method irrespective of approach to tackle these two Challenges are difficult. Based on the distortion method used, a unique approach is needed for measuring privacy and accuracy.

One open problem with respect to privacy-preserving data is variable privacy levels. A number of surveys conducted on users have led to the conclusion that multiple privacy levels are needed [9]. The rationale being, each user may have a different Level of sensitivity when answering a question or presenting some data. If the user is not satisfied with the privacy level offered by the system, the user may not answer the question. Offering the user with multiple levels of privacy would help the user to have discretion when needed without the system having to compromise on accuracy where unnecessary.

1.3 MULTIPLE PRIVACY LEVELS

Let us consider a super-market billing database, with ten items namely; Bread, Cheese, Beer, Diaper, Medicines, Toys, Biscuits, Stationary, Water, and Magazines. Say there exist data about 4 users. The data consists of items the user has bought on a single bill. The example data is as shown in table 1.1.

Table 1.1: Sample Transaction 1

User Items	Items
User1	Bread , Medicines ,Beer ,Toys ,Cheese ,stationary
User2	Stationary ,Toys ,Beer ,Bread , Cheese ,Medicines ,Diaper
User3	Toys ,Beer ,Bread , Medicines ,Diaper
User4	Stationary ,Toys ,Beer ,Bread ,Medicines ,Diaper

Let us assume a randomization method that replaces two items from each existing transaction with two new items. Let the randomized transaction be as shown in table 1.2

Table 1.2: Sample Transaction 2

User Items	Items
User1	Bread , Medicines ,Beer ,Toys ,Water ,Magazines
User2	Stationary ,Toys ,Beer , Water, Magazines ,Medicines ,Diaper
User3	Toys ,Beer ,Bread , Medicines ,Diaper
User4	Water, Toys ,Beer ,Bread ,Medicines ,Diaper

The parameter in the above randomization method is the number of items that are replaced. Let that parameter be R. Here R = 2. All transactions in the above set of transactions are randomized with the same value for parameter R, the higher the value of R, the higher the privacy. However, for a high value of R, loss of accuracy is greater. For example: For $R \geq 4$, almost all items from the original transactions are replaced by random items. It may give maximum privacy but no utility. Optimal value for the randomization parameter is paramount in randomization-based systems.

Say users User2 and User3 are giving data to the above system. User2 is a more conservative person and requests high privacy. Whereas, User3 is more relaxed and wishes to support the experiment and does not worry about privacy. There are two options for the value of randomization parameter. It can be set high and lose accuracy or it can be set low and some of the users may not provide any data. Lack of data will equally affect the accuracy of the system. In privacy based systems there can be a compromise on accuracy but not on privacy. One solution recommended for this problem is to use multiple privacy levels. Instead of using one parameter to randomize all transaction equally, the user can choose their privacy level depending on their privacy requirement.

Say the users choose the following values;

User1 R = 0

User2 R = 1

User3 R = 2

User4 R = 3

Then the randomized transaction may look like;

Table 1.3: Sample Transaction 3

User Items	Items
User1	Bread , Medicines ,Beer ,Toys ,Water ,Magazines
User2	Stationary ,Toys ,Beer , Water, Magazines ,Medicines ,Diaper
User3	Stationary ,Beer ,Bread , Magazines ,Diaper
User4	Water, Toys ,Beer ,Bread ,Medicines , Medicines, Cheese

By using multiple privacy levels, the system can achieve a reasonable trade-off between privacy and accuracy.

1.4 APRIORI ALGORITHM

Apriori is a commonly used algorithm for finding frequent item sets from a set of transactions [2]. It uses a bottom up approach to find frequent subsets. These frequent subsets are extended one item at a time and they are tested against the data.

The algorithm terminates when no successful extension happens in iteration. The frequent item sets determined by Apriori are used to determine association rules. It has applications in domains such as market basket analysis. Market Basket analysis is the application of association rule mining in retail industry. It helps retailers to problem customers based on purchasing pattern. This information can be used to improve their sales. For example, an urban legend suggests market basket analysis on customers showed that beer and diaper are often purchased together, so putting the items close to each other can improve sales of the two products. Association rule mining finds applications in fields such as; medical research, Criminal behaviour, counter-terrorism, etc.

II. LITERATURE SURVEY

In privacy-preserving data mining, many methods are proposed, each with a different situation where it can be used. The system considered in this paper is based on an approach where the data is incremental. The mining activity is conducted at a single location/server. In this system, the users need not depend on third party or central server for privacy guarantees. For these requirements, a randomization-based system for privacy preservation is appropriate.

Randomization is one of the methods used for privacy preservation in data mining. It is widely used for systems where data need not be stored in its original form (Example: Online survey, data for advertising, etc.). This method is useful when the exact data mining operation is not known. The mining algorithm, expected results, and attributes needed are not decided beforehand. It guards from malicious data providers and is an open system (Trust between parties not an absolute requirement). However, it suffers from some well-known disadvantages such as; spectral analysis attack [4] (even after randomization, the relationship between the attributes are retained) and public data attack (if part of the data is available in public domain, it can be used to compensate privacy). Various other methods of privacy preservation are discussed in this chapter. In randomization, we do not store data as is; instead the data can be randomized using some pre-defined method and this randomized data can be stored [3]. The original records cannot be recovered but certain characteristics of the original data can be estimated using the randomized data [3]. Thus, privacy can be understood as; the difficulty of estimating the original data. Utility can be understood as the (comparatively lower) difficulty of estimating the distribution of original data.

2.1 PRIVACY PRESERVATION METHODS

Privacy preservation methods can be classified under three categories:

- Data collection
- Inference control
- Information sharing

The different methods in privacy preservation

2.1.1 DATA COLLECTION

Data collection deals with the format in which data is collected and stored.

There are three types of data collection methods.

- Data Exchange
- Noise Insertion
- Cryptography

2.1.1.1 DATA EXCHANGE

In data exchange, a value from a selected attribute and tuple is exchanged with another tuple of the same attribute. By this method, the distribution of the attribute is left untouched. But, an attribute value of a selected tuple cannot be given with high accuracy. A trusted server is needed to get the original value before exchange.

There is also a risk with malicious data providers. They can give values outside valid values, which will increase the probability of compromising privacy. The user has to depend on the entry of at least one other person for ensuring their privacy.

2.1.1.2 NOISE INSERTION

Noise insertion is one of the widely used methods. Noise insertion is also known as randomization. Data perturbation is one such method of randomization proposed in [3]. In data perturbation, randomized noise is inserted into the data (randomized noise is generated either using an uniform or Gaussian distribution). It can be used for classification algorithms such as decision tree classifier and naive Bayesian classifier. After perturbation of data, a reconstruction algorithm is used to get back the original distribution. The original data values cannot be reconstructed.

In the context of association rule mining, Warner's randomized response is used implicitly [5]. Data in association rule mining is in the form of transactions. Each trans action ($t_1; t_2; t_3; t_n$) is a subset of items from the item set space $I(i_1; i_2;$

$i_3; \dots; i_m$). In most methods, from a given transaction t_x with k items being present in that transaction; l (where, $l < k$) elements are chosen at random with a pre-defined probability p .

And the remaining items from the item set that are not part of the original transaction are selected with a probability q . The selected items are said to be part of the new randomized transaction $t_0 x$. Similarly the transactions are perturbed and frequent item set are derived. Using the randomized support and the randomization function, original support of the chosen item set is estimated [5]. Randomization methods for association rules are of two types; Using randomized response method [5] and using matrix based methods [3]. Randomized response based methods are efficient but privacy reduces drastically for item set size > 3 . Item set size > 3 is left by the author as an open problem [5]. For matrix based methods, higher item set size does not affect privacy but the process is time consuming. Because, in each transaction the entire item space is stored in a matrix [3].

2.1.1.3 CRYPTOGRAPHY

Cryptography based methods are similar to noise insertion. Cryptographic methods are used on the original data and the encrypted data is stored and used directly to estimate data mining results without reconstructing the original data. In this method, the number of users who can enter data to the system is limited. Users who wish to enter data to the system have to be given a key pair to encrypt their data. The process of setting up systems that need data from a large number of sources is difficult.

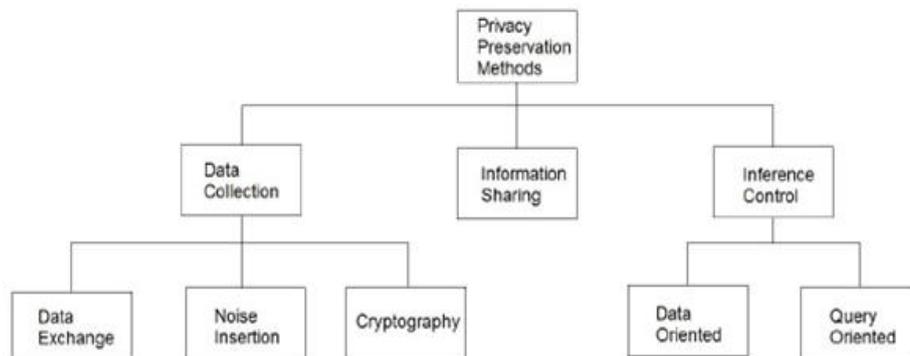


Figure 2.1: PRIVACY-PRESERVING DATA MINING METHODS

2.1.2 INFERENCE CONTROL

Inference control methods are of two types;

- Data oriented
- Query oriented

2.1.2.1 DATA ORIENTED

In data oriented methods, sensitive data are identified and their values are distorted during query response or during storage.

2.1.2.2 QUERY ORIENTED

In this method, a safe set of queries is either pre-determined or calculated during runtime. And any query within this safe set is replied to and others are ignored. This makes the system less available. Efficiency is affected if the system has to compute safe set of queries at runtime.

2.1.3 INFORMATION SHARING

All methods of information sharing are variations of Secure Multi-party computation. Each user depends on at least one other user. A user sends his data to another user, who computes aggregation functions before sending it to the server. It has two restrictions; Semi-Honest and Malicious restriction. Due to these restrictions, privacy will be compromised if malicious users are introduced in the system.

2.2 K-ANONYMITY

K-Anonymity is a method which guarantees that, under any set of selection conditions, at least K rows will be selected. In other words, ability to uniquely identify rows is minimized. This prevents the unique identification and mapping of a single transaction to a specific data source/owner. Let $A(A_1; A_2; A_3; \dots)$ be some set attributes. For some set of values $(a_1; a_2; a_3; \dots)$ given as $(A_1 = a_1; A_2 = a_2; A_3 = a_3; \dots)$, there exists at least K tuples that are selected.

3.3 WARNER'S RANDOMIZED RESPONSE

It is a method used for collecting private information in surveys. It was proposed in 1965 [7]. The method is based on uniform randomization. In a survey, a 'YES'/'NO' question is asked. Before answering, a coin is flipped. If the coin comes up as tail, the answer is always recorded as 'YES'. And if the coin comes up as head, the actual answer given is recorded. The coin toss can be done by the person answering the survey, thus only that person may know the real answer. There is 'NO' need for a trusted third party.

Half the answers get tails and the other half get heads when they flipped the coin. Therefore, half of those answers which were \NO", will be recorded as \YES". So the actual number of \NO" answers will be double the recorded number. For example, if 20% of the answers are \NO", then the true fraction that replied \NO" is 40%.

III. PROPOSED METHODOLOGY

3.1 TERMS AND DEFINITIONS

3.1.1 SUPPORT

Support is a metric used for finding interestingness of rules that have been identified in association rule mining. It can be defined as the ratio of number of times an item set occurs over the total number of transactions. The higher the support, the more frequent the item set. There are other interestingness measures such as lift and confidence.

3.1.2 CONFIDENCE

Confidence is another metric used to find interestingness of rules that have been identified in association rule mining. Given a rule $X \Rightarrow Y$, Where X and Y are subsets of all unique items from the given transactions. It can be defined as the ratio of number of times X and Y occur together over the total number of times X occurs in all transactions.

3.1.3 PRIVACY BREACH

To measure the level of privacy obtained in our system, a metric known as privacy breach [5] is used. Given that an item set occurs in a randomized transaction, the probability of its occurrence in the original transaction is calculated as the privacy breach level. We measure privacy breach as: Given a frequent item set appears in a randomized transaction, the number of times it appears in the corresponding original non-randomized transaction [5]. The lower the privacy breach level, the higher the privacy.

3.1.4 ACCURACY

To measure the utility of the system, a comparison is made between the rules that have been identified in the randomized transactions and those that have been identified in the original transactions. A tabular column is made to identify true rules, true positives, false positives and false drops.

3.2 PROPOSED ARCHITECTURE

We focus on a method such that the original data is not stored in any central server and the user can themselves randomize the data before sending to the mining server. In this way, the user does not need to trust the central server. There are multiple steps in data mining. First, the data is collected from their sources. An online survey system example can be taken, where the survey respondents are the data source or Data owners.

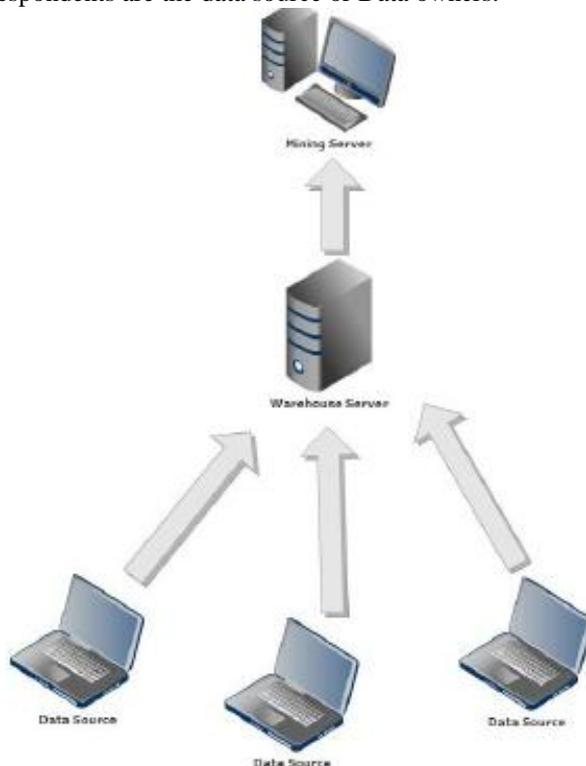


Figure 3.1: Architecture Diagram

They provide data to the data warehouse server. There can be more than one source. The data sources form the lower level of the architecture. The data mining server forms the higher level of the architecture. There can be a warehouse server in between to store and process the data for mining. The sample architecture is shown in shown in figure 1.1.

3.3 METHODOLOGY

In this paper three methods are implemented for our experiments. The process of each experiment can be broadly explained using the chart shown in figure 4.2. The original dataset is obtained, prepared and reduced. Existing apriori method to run frequent item sets is executed with parameters set for support and minimum item set size. One of the randomization methods are run on this original dataset to obtain a randomized dataset. Apriori algorithm is run on this dataset to obtain frequent item sets. The parameter for minimum item set size is left unchanged. The support value is reduced slightly following our assumption from previous chapters that support of frequent item sets may reduce due to randomization. And the probability of increase of support is very less. As a baseline for comparison, uniform randomization with single level of privacy is implemented. Privacy and utility metric are calculated for this method. Uniform privacy with multiple privacy levels is implemented and compared with single privacy method. Finally, weighted randomization method is implemented and the privacy and accuracy are calculated, compared and tabulated. The parameters used in the following methods are shown in table 1.2

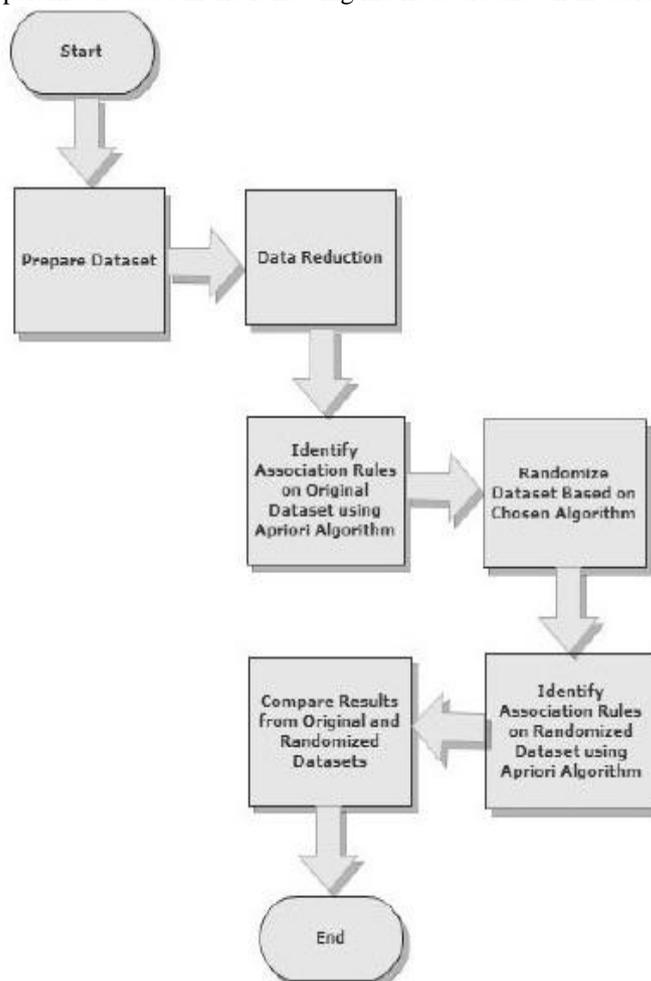


Figure 3.2: Flow Diagram

Table 3.1: Parameters

Parameter	Meaning
Lmax	Total number of privacy levels in the system.
Ki	Cut-parameter of the ith level. For single level privacy, $i = 1$. ($1 \leq i \leq lmax$)
Pi	Randomization level of the ith privacy level. For single level privacy, $i = 1$. ($1 \leq i \leq lmax$)
L	Limit parameter, to limit the size of randomized transaction. This parameter is used to reduce the dimensionality of randomized transaction.
T	Set of all transactions.
N	Total number of transactions in T.
M	Total number of items in I.
Tj	The jth transaction selected from T. ($1 \leq j \leq N$)
tj	The transaction tj after randomization. ($1 \leq j \leq N$)
I	List of all unique items.

3.3.1 UNIFORM RANDOMIZATION

With Single Level A uniform randomization method is implemented. It follows Warner's randomized response method. A cut-o_ value k , and randomization value p are selected. For a given transaction, k items are selected and randomization is performed by adding $L \square k$ items from the list of all unique items. Each item from the unique list of items can be selected with a probability of p . The list of items thus obtained is taken as $t_0 j$.

The method is explained in below.

- Step 1: Select a cut-o_ value k , and randomization value p .
- Step 2: Get the list of all unique items from all transactions.
- Step 3: Read transactions one by one from all transactions T .
- Step 4: Let the current transaction is t randomly select k items from the current transaction t and add them to the randomized transaction t_0 .
- Step 5: From the list of all unique items, select L items randomly; each with a probability p . L is a limiting parameter on the transaction size of randomized transactions.
- Step 6: These L items selected randomly, should not be present in original transaction t . They are added to t_0 .
- Step 7: Repeat the same operation on all such transactions.
- Step 8: Run association rule mining on the set of randomized transactions obtained.

3.3.2 UNIFORM RANDOMIZATION WITH MULTIPLE LEVELS

In this method, Cut-o_ values K_i , and randomization value p_i are selected for each of the levels. For a given transaction and privacy level, K_i items are selected and randomization is performed by adding $L \square k_i$ items from the list of all unique items. Each item from the unique list of items can be selected with a probability of p_i . The list of items thus obtained is taken as $t_0 j$. The method is explained below.

- Step 1: For each of i levels, Select cut-o_ values K_i , and randomization values p_i .
- Step 2: Get the list of all unique items from all transactions.
- Step 3: Read transactions one by one from all transactions T .
- Step 4: Let the current transaction is t . Select one of the i randomization levels. Let the selected level cut-o_ value be k and randomization level is p .
- Step 5: Randomly select k items from the current transaction and add them to the randomized transaction t_0 .
- Step 6: From the list of all unique items, select L items randomly; each with a probability p . L is a limiting parameter on the transaction size of randomized transactions.
- Step 7: These L items selected randomly, should not be present in original transaction t . They are added to t_0 .
- Step 8: Repeat the same operation on all such transactions.
- Step 9: Run association rule mining on the set of randomized transactions obtained.

3.3.3 WEIGHTED RANDOMIZATION

In the previous two methods, the frequency of occurrence of items encountered is not considered. Due to the cut-o_ parameter, some items from the original transaction are removed. In this process, some frequent item sets may be removed. To balance this during randomization rather than adding all items with equal probability.

Items are added with a probability that is directly proportional to their frequency of occurrence. Thereby, high frequency items occur often and low frequency items occur rarely. This process may increase the support of frequent item sets.

- Step 1: Select a cut-o_ value k .
- Step 2: Read transactions one by one from all transactions T .
- Step 3: Create a data structure that stores all unique items that have been encountered.
- Step 4: Let the current transaction be t_i .
- Step 5: Maintain in the data structure all unique items that have been encountered and the number of times each has occurred, in all transactions before the current transaction ($t_0; t_1; t_i \square 1$).
- Step 6: Till the transaction t_i , say there were j unique items. Then let the unique set of items in the data structure be, $item_1; item_2; \dots; item_j$. And each with its corresponding number of occurrences as, $c_1; c_2; \dots; c_j$.
- Step 7: Randomly select k items from the current transaction and add them to the randomized transaction t_0 .
- Step 8: From the list of all unique items encountered, select each item (item X) randomly, with a probability $c_x = i$.
- Step 9: The items selected randomly, should not be present in original transaction t they are added to t_0 .
- Step 10: Repeat the same operation on all such transactions.
- Step 11: Run association rule mining on the set of randomized transactions obtained.

IV. DATASET

The experiments are carried out on one real-life dataset. The WorldCup98 dataset is generated from all the requests made to the 1998 World Cup website between April 30, 1998 and July 26, 1998. During this period of time the site received 1,352,804,107 re-quests. It is publicly available at <ftp://research.smp2.cc.vt.edu/pub/worldcup/> 4. The scanned log is used to produce a transaction, 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.

The format of web requests is;

```
struct request
{
```

```
uint32_t timestamp; uint32_t
clientID; uint32_t objectID;
uint32_t size; uint8_t method;
uint8_t status; uint8_t type;
uint8_t server;
```

The Fields of the request structure contain the information as shown in table 5.1.

Parameter	Meaning
timestamp	The time of the request, stored as the number of seconds. The timestamp is in GMT
client ID	A unique integer identifier for the client that issued the request. Due to privacy concerns these mappings are not released. Each client ID maps to exactly one IP address, and the mappings are preserved across the entire data set - that is if IP address 0.0.0.0 mapped to client ID X on day Y then any request in any of the data sets containing client ID X also came from IP address 0.0.0.0 object ID- A unique integer identifier for the requested URL. These mappings are also 1-to-1 and are preserved across the entire data set
size	The number of bytes in the response
Method	The method contained in the clients request (e.g., GET)
status	This Field contains two pieces of information. The 2 highest order bits contain the HTTP version indicated in the clients request (e.g., HTTP/1.0). the remaining 6 bits indicate the response status code (e.g., 200 OK)
type	The type of requested (e.g., HTML, IMAGE, etc.), generally based on the file extension (.html), or the presence of a parameter list (e.g.,? Indicates a DYNAMIC request). If the url ends with, it is considered a DIRECTORY
server	Indicates which server handled the request. The upper 3 bits indicate which region the server was at (e.g., SANTA CLARA, PLANO,HERNDON, PARIS). the remaining bits indicate which server at the site handled the request. All 8 bits can also be used to determine a unique server

To qualify to be an item, the request must satisfy the following: 1. Clients request method is GET. Request status is 200.

File extension is .htm or .html.

Items of same session are considered as one transaction.

Items do not correspond to the French and English front page frames. The les /English/frntpage.htm and /French/frntpage.htm are removed as every user has to access the FrontPage request to access the website. This page was accessed in every session and is trivial.

A session consists of requests from only one client, Starts with a request that satires the above properties, and ends when the last click from this Client ID has exceeded 30 minutes. All empty transactions and all transactions of size above 10 are deleted. The dataset is further reduced by considering items from the last 19 days between July 8, 1998 and July 26, 1998. The resulting WorldCup98 dataset consists of 1,214,174 transactions, and 6,708 items. The dataset used in this thesis consists of a set of items accessed per session by a user. Finding association rules on this dataset can help to improve performance of a tool that suggests content to users. Association rules tell which les users are more inclined to accessing given they have already accessed a set of items. It is similar to market basket dataset, the analogy being users are inclined to buy a set of items together. Consider a survey where users are requested to give the sites they visited during the day, rather than the items they accessed from a given HTTP site. Some users may hesitate as they may feel embarrassed to give all the sites they visited, due to various reasons. This may the accuracy of the system. Due to randomization, some not-so-popular sites may get removed from certain transactions and would get added to some other transactions. The data miner knows of such randomization. Therefore, the data miner cannot be 100% sure if the user has really accessed the site or not. This uncertainty is the privacy the user gets.

4.1 METHOD FOR DATASET PREPARATION

Step 1: Convert the format of data from binary log to common log using the tools (ftp://ita.ee.lbl.gov/software/WorldCup_tools.tar.gz), given along with the dataset.

Step 2: The dataset is partitioned based on days. Open each le, and read line by line.

Step 3: Check if the selected line is a get request, reply status was success, the le eld is not empty, the le is not FrontPage and the le is a html le.

Step 4: In a separate data structure, maintain client id, last request time and items (le names) accessed (in chronological order, separated by commas).

Step 5: In this data structure, check if current client id exists and deference between access time in data structure and current client id is not more than 30 minutes. Step 6: If step 5 is true, update access time in data structure with current access time, append the current item (le name).

Step 7: If step 5 is false, write transaction list (items accessed) to a le that has all processed transactions, and update access time in data structure with current access time, delete all items in item list and set items accessed as the current item (le name).

Step 8: When the le is completely processed, write all client ids in data structure to the le containing processed transactions.

Step 9: Maintain a counter that counts number of transactions written to While writing the processed transactions to a le, write in the format of (transaction number, item), and each item in a separate line.

The WorldCup98 dataset is prepared using this method. It is an open dataset. The format used in WorldCup98 dataset makes it easier to convert it into transactions. It has already been used in association rule mining experiments.

V. RESULTS

The WorldCup98 dataset consists of items accessed from world cup 98 website. Three randomized datasets are generated from the WorldCup98 dataset. Frequent item sets are indented from these data using R [6]. The rest randomized dataset is generated by randomizing the WorldCup98 dataset using uniform randomization with single privacy level method. Unique list of items from all transactions is obtained by parsing the list of transactions once. Then each transaction in the WorldCup98 dataset is read. Cut-parameter number of items is added to the randomized transaction. From the list of unique items, an item which is not part of the original transaction is selected at random. With a probability depending on the probability parameter chosen, the selected item is added to the randomization transaction. Similarly items are added to the randomized transaction until it contains limit parameter number of items. The item set thus generated is the SingleWorldCup98 dataset.

The second randomized dataset is generated by randomizing the WorldCup98 dataset using uniform randomization with multiple privacy levels method. Two privacy levels are chosen in this method. Unique list of items from all transactions is obtained by parsing the list of transactions once. Then each transaction in the WorldCup98 dataset is read, a privacy level is chosen at random for every transaction. Cut-o parameter and probability parameter vary depending on the level of privacy selected. Cut-parameter number of items is added to the randomized transaction. From the list of unique items, an item which is not part of the original transaction is selected at random. With a probability depending on the probability parameter chosen, the selected item is added to the randomization transaction. Similarly items are added to the randomized transaction until it contains limit parameter number of items. The item set thus generated is the MultipleWorldCup98 dataset, the third randomized dataset is generated by randomizing the WorldCup98 dataset using weighted randomization method. Each transaction in the WorldCup98 dataset is read. The items form this transaction are updated in a separate table. In this table, all items that have been encountered till that transaction and the number of times each of them occurred are recorded. Cut-parameter number of items is added to the randomized transaction. From the table that stores the list of items encountered, each item is added to the randomized transaction with a probability of the ratio between frequencies of occurrence over total number of transactions read (frequency/number-of-transactions). The item set thus generated is the Weighted-WorldCup98 dataset.

The WorldCup98, SingleWorldCup98, MultipleWorldCup98 and WeightedWorldCup98 datasets are loaded into R and frequent item sets are generated using Apriori algorithm from rules package. The minimum support used for WorldCup98 dataset is 0.2. The SingleWorldCup98 and MultipleWorldCup98 datasets are run with a minimum support of 0.15. The WeightedWorldCup98 dataset is run with a minimum support of 0.2.

The results presented are tabulated under the following headings; true item sets, true positive, false drop, false positive. True item sets are frequent item sets that are indented from the original non-randomized transaction namely WorldCup98 dataset. True positives are frequent item sets that are indented from the randomized dataset which are part of the frequent item sets indented from WorldCup98 dataset. False drop are frequent item sets that are indented from WorldCup98 dataset but are not indented from the randomized dataset. False positives are frequent item sets that are indented in the randomized dataset but are not indented from the WorldCup98 dataset. These results are presented based on item set size of the frequent item sets indented. Separate tables are presented for each of the randomized datasets namely Single-WorldCup98, MultipleWorldCup98 and WeightedWorldCup98 shows the parameters used for randomization with single privacy level experiment. 6.2 show the comparison between WorldCup98 dataset results and those obtained using the uniform randomization method with a single level.

VI. CONCLUSION

The work presented above show that without modifying the original algorithms for association rule mining it is possible to get accurate results. The privacy breach level calculated is high for Uniform randomization with single level. With multiple levels and in weighted randomization methods, the privacy breach level reaches an acceptable level of 50% or less. In this paper, the major contributions are; to show that multiple levels of privacy can be implemented without lowering the accuracy greatly. Multiple levels also decreased the privacy breach level slightly (Increased privacy). Weighted randomization is based on all the transactions that have been encountered. In this method, the privacy breach level falls further below the uniform randomization methods.

VII. FUTURE WORK

The minimum support used in the previous methods is arrived at by experimentation. The values used for minimum support may not be optimal in each case. Finding minimum support to get a trade-o_ between accuracy and privacy is not trivial, and is left as an open problem. We have shown that the weighted randomization method gives better privacy than

uniform randomization. Multiple privacy levels do not affect accuracy by a large amount at the same time, giving an increase in privacy level. Implementation of multiple privacy levels for weighted randomization is left as a future area for research.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my supervisor Assistant prof Vajiy Kumar for his constant guidance and support throughout the paper without his support this paper would have not been completed , I would also thank all the CSE department teachers for providing me opportunity to pursue research , I thank my family . the constant inspiration and guidance kept me focus and motivated and thanks Chandigarh University .

REFERENCES

- [1] C. C. Aggarwal and P. S. Yu, editors. Privacy-Preserving Data Mining – Models and Algorithms, volume 34 of Advances in Database Systems. Springer, 2008.
- [2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. of 20th Intl. Conf. on VLDB, pages 487{499, 1994.
- [3]] R. Agrawal and R. Srikant. Privacy-preserving data mining. SIGMOD Rec., 29(2):439{450, May 2000.
- [4] A. Ev_mievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In Proceedings of the twenty-second ACM SIGMOD- SIGACT-SIGART symposium on Principles of database systems, PODS '03, pages 211{222, New York, NY, USA, 2003. ACM.
- [5] A. Ev_mievski, R. Srikant, R. Agarwal, and J. Gehrke. Privacy preserving mining of association rules. Inf. Syst., 29(4):343{364, June 2004.
- [6] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.
- [7] Wikipedia. Warner's randomized response, 2004.
- [8] Q. YANG and X. WU. 10 challenging problems in data mining research. International Journal of Information Technology and Decision Making, 05(04):597604,2006
- [9] N. Zhang. Privacy preserving data mining: Ph.d dissertation, 2006.