Asian Journal of Information Technology 15 (2): 298-311, 2016
ISSN: 1682-3915

Proposing New Strategy for Privacy Preserving Microdata Publishing with Conditional Functional Dependencies

[1]S. Balamurugan and [2]P. Vislakshi
[1]Department of Information Technology, KIT-Kalaignarkarunanidhi Institute of Technology,
[2]Department of Electronics and Communication Engineering, PSG College of Technology,
Anna University Chennai, Coimbatore, Tamil Nadu, India

Abstract: Publishing individual data has grabbed more attention towards individual privacy in recent days. It has become an active research area and several works have been carried out to preserve privacy. The (d, l) Inference Model each and every group contains l sensitive values with frequency similarity value and which is controlled by the parameter d. Recent researches do not focus on CFD against the (d, l) Inference Model and doesn't work with a Frequency Distribution Method for initial partition phase. In order to overcome these two problems, this study proposes a novel (d, l) Inference Model to deal with adversarial information rendered by Conditional Functional Dependencies (CFDs). However, discovering the quality of CFD is a challenging task. In order to solve the above mentioned problem, Automatic Compact Frequent Pattern Growth Branch Sort algorithm (CFPGBS) is proposed for mining the best CFD patterns and removing the less quality CFD patterns. A compact pattern tree is developed, that captures CFD patterns information with insertion phase and provides the better pattern mining performance for CFD patterns. The construction of the initial partitions for the bottom-up approach driven is performed by Log-Skew-Normal Alpha-Power distribution (LSKNAPD) frequency distribution function. Experimental results show that the proposed (d, l) Inference Model can proficiently anonymize the micro data with a low information loss against CFD.

Key words: Conditional Functional Dependency (CFD), privacy preservation, Functional Dependency (FD), bottom-up approach, compact pattern tree, frequency pattern growth tree, skew normal frequency distribution

INTRODUCTION

Publishing micro data within the organizations on government and non governmental agencies generates which in turn makes preserving privacy of micro data an essential task. These data can unintentionally disclose perceptive personal data to malicious adversaries. Hiding individual attribute information such as name and race does not be satisfactory to protect privacy. The intruder may still easily determine the authenticated information, by combining the different micro data table attributes released results are publicly existing data. Generally, these types of the attributes are categorized into two types, namely Quasi-Identifiers (QIs) and Sensitive Attributes (SA). Table 1 shows the example of micro data table with quasi identifiers (QIs) and Sensitive Attributes (SA).

In order to solve the problem of the privacy preserving, several anonymization and suppression techniques have been developed; still the existing anonymization techniques leak individual sensitive attribute information during certain attacks. Past research works used syntactic and perturbation-based methods to ensure privacy assurance of anonymized data. A syntactic anonymization technique assumes that all attributes in the micro table are equally partitioned into set of Equivalence Classes (ECs), in such, a way that every attribute in the micro table inside a ECs are mutually identical with every others attributes in the Microtable as distant as their QIs are disturbed (Xiao *et al*., 2010).

The working principle of these models differs based on the ECs. By k-anonymity, every EC must be incorporated as a minimum of multi iterative k tuples (Chhinkaniwala and Garg, 2014). In k anonymization methods, the privacy for QI attributes are only protected, but it does not concentrate on the values of a SA; therefore, the privacy concerning such values may possibly be integrated. In order to deal with the problem of k anonymization, l-diversity methods that contain at

Corresponding Author: S. Balamurugan, Department of Information Technology,
KIT-Kalaignarkarunanidhi Institute of Technology, Affiliated to Anna University Chennai, Coimbatore,
Tamil Nadu, India

Table 1: A micro data table of hospital patient discharge data

Quasi Identifier (QI)						Sensitive Attributes (SA)	
CO	AC	H	A	G	ZC	R	ICD-9-CM
India (01)	022	111111	39	F	71000	Asian	HIV
India (01)	022	111111	41	M	72001	White	Diabetes
America (02)	45	222222	58	F	73021	Black	Diabetes
India (01)	024	222222	43	F	72001	Black	Flu
India (01)	024	111111	48	M	73020	Black	Alcoholism
America (02)	45	111111	51	F	71000	White	Diabetes
India (01)	023	333333	56	M	72000	White	Flu

least l different "well characterized" SA values for each Quasi Identifiers can be used (Wang and Wang, 2013).

But still, l-diversity methods are not successful to protect personal information from attacker for every SA value's frequency in a released table. As an adaptation to this issue, a t-closeness condition with distribution function has been introduced. This method determines the frequency distribution between SA values and original distribution result. Yet, t-closeness condition fails to protect privacy fully functional dependency functions for effectual and human-understandable policy (Nergiz *et al*., 2007). Moreover, recently, k anonymity and l-diversity techniques have been improved through t-closeness, (α, k) anonymity (Wong *et al*., 2006) and (c, k) safety (Martin *et al*., 2007) approaches. All these privacy preservation models overcome the issues of different adversary information. But none of the above mentioned works follow the Full Functional Dependencies (FFDs) based privacy protection.

If the attacker identifies FFDs adversary information even after applying the privacy preserving principles, the intruder can easily violate the principles of privacy. But the major problem of FFD (Wang and Liu, 2011) is that it does not measure the semantically related constants between attributes among the data while performing (d, l) Inference Model. The main focus of this study is to measure the semantic relationship among the attributes with CFD to discover data validation rules for (d, l) Inference Model privacy. Recently, CFDs have been developed to discover data inconsistency which is the extension of Functional Dependencies (FD) and provides a framework dependent solution by SQL (Medina and Nourine, 2009; Cormode *et al*., 2009). The major contributions of the research as follows:

- Analyzing the differences in Conditional Functional Dependencies (CFDs) and Fully Functional Dependencies (FFDs) considered as adversarial knowledge
- Proposing a (d, l) Inference Model for privacy protection for CFD. It requires l well represented SA that are of similar frequency, where the correspondence of the inference model is controlled by frequency distribution d with CDF. Mining the CFD for (d, l) Inference Model using compact frequent pattern growth branch sort algorithm
- Development of a well-organized anonymization algorithm that generates the (d, l) inference anonymization system with low information loss. It consists of two major steps, phase-1 partition and phase-2 QI-group construction. In order to perform this phase-1 partition, the construction of the initial partitions for the bottom-up approach has to be driven by the Log-Skew-Normal Alpha-Power distribution. Thus, bottom-up approach is presented to make partitions with less information loss. Three model strategies is be referred from (Wang and Liu, 2011) to group sensitive values results from Log-Skew-Normal Alpha-Power distribution into smaller partitions to decrease information loss and evaluate with other methods
- Finally, this work shows the efficiency of the proposed CFDs (d, l) Inference Model interms of reduced time and information loss by wide range of simulations

Importance of conditional functional dependency than existing functional dependency

Functional Dependency (FD): A Functional Dependency (FD) is a category of integrity restriction. Two attributes X and Y are specified and a database instance D should satisfy the FD for the condition $F:X\rightarrow Y$ if for each two tuples in database D t_1, $t_2 \in D$, if $t_1.X = t_2.X$ then $t_1.Y = t_2.Y$. Then, it is called as determinant attributes and dependent attributes for X and Y and their tuples values t_1 and t_2 correspond to the determinant values and dependent values of X and Y. FD $F:X\rightarrow Y$ is named as Fully FD (FFD) if it holds for every values of two attributes X and Y. Otherwise, it is called a Conditional FD (CFD). In this study, both FFD and CFD are considered for privacy preserving and the major contribution of the research focus on CFD for publishing microdata table.

Let D be database instance microdata table that stores the confidential information of a set of persons. Database D consists of two attributes: Quasi-Identifiers QI, whose combination can take part in individual key of the patient discharge data as represented in Table 1. The patient discharge data relation can be related to customers through attributes such as Country (CO), Area Code (AC), Hospital (H), Age (A), Gender (G), Zipcode (ZC), Race (R) and Disease (D) where Race (R) and ICD-9-CM corresponds to Sensitive Attributes (SA) and other attributes corresponds to Quasi Identifiers (QI). In order to appreciate the idea of FFD and overcome the limitations of FFD, a framework of CFD with (d, l) Inference Model is used in the present research work. It is assumed that Table 1 data includes the functional dependency F:Race→Zip which states that any two similar races should match to the similar zipcode.

It is to be assumed that the attacker possesses the information of F. Then, while applying Functional Dependency F on the 3-diversity table in Table 2, since the second group contains the races Black and white with zipcode "72***" the attacker can modify the zipcode value of the second and fourth tuples "7200*" to "72***" In order to overcome this problem, FD is applied to anonymized result as shown in Table 3. By the verification relation attack, the privacy promise on the first tuple simply satisfies only 1-diversity. Thus, by the verification relation attack, the attacker can explicitly decide that an Indian has disease Diabetes with the zip code "7200*".

In order to overcome this problem, FFD is considered by taking frequency or number of occurrences in the each and every dependency function. D-closeness (distribution) based closness to address the constraint of closest frequency. In particular, two data values s_1 and s_2 are declared as closest distribution d-close if $|f_1\text{-}f_2|\leq d$ where f_1 and f_2 are the count occurrence of s_1 and s_2. But in this full functional dependency, the count frequency of the Sensitive Attributes (SA) does not regard as the semantic association amongst the attributes. Traditional FFDs that hold on relation r_0 include the following:

$$f_1\text{: } [\text{CO, ZC}] \rightarrow \text{R} \tag{1}$$

$$f_2\text{:}[\text{CO,ZC,H}] \rightarrow \text{disease} \tag{2}$$

Here, f_1:[CO, ZC]→R implies that two persons with the same country (India or America) and area code should also have the same city (India or America).

This process is same for f_2. Since, the probability of every attribute to overcome this problem is lesser, CFD is introduced that can hold on r_0 which consists of not only the FFDs f_1 and f_2 but also the following:

$$\phi_0\text{: } \left([\text{CO, ZC}] \rightarrow \text{Disease}\left(01,\_\ ||\ \_\right)\right)$$
$$\phi_1\text{: } \left([\text{CO,H}] \rightarrow \text{Disease}\left(01,023\ ||\ \text{Diabetes}\right)\right)$$
$$\phi_2\text{: } \left([\text{CO,H}] \rightarrow \text{R}\left(02,40\ ||\ \text{black}\right)\right)$$
$$\phi_3\text{: } \left([\text{CO,H}] \rightarrow \text{Disease}\left(01,022\ ||\ \text{HIV}\right)\right)$$

In ϕ_0, disease (01,_||_) is the pattern tuple that enforces a building of semantically related constants for attributes CO, ZC, Disease in a tuple. It states that for customers in the India, ZC uniquely determines disease. It is FFD that only holds on the subset of tuples with the pattern "CO = 01", rather than on the entire relation r_0.

Table 2: A microdata table before FD of hospital patient discharge data

Quasi Identifier (QI)						Sensitive Attributes (SA)	
CO	AC	H	A	G	ZC	R	Disease (D)
India (01)	022	111111	37-45	*	7****	Asian	HIV
India (01)	022	111111	37-45	*	7****	White	Diabetes
America (02)	45	222222	48-56	*	7****	Black	Diabetes
India (01)	024	222222	37-45	*	7****	Black	Flu
India (01)	024	111111	48-56	*	7****	Black	Alcoholism
America (02)	45	111111	48-56	*	7****	White	Diabetes
India (01)	023	333333	48-56	*	7****	White	Flu

Table 3: A microdata table After FD of hospital patient discharge data

Quasi identifier(QI)						Sensitive attributes (SA)	
CO	AC	H	A	G	ZC	R	ICD-9-CM
India (01)	022	111111	37-45	*	71***	Asian	HIV
India (01)	022	111111	37-45	*	72***	White	Diabetes
America (02)	45	222222	48-56	*	73***	Black	Diabetes
India (01)	024	222222	37-45	*	72***	Black	Flu
India (01)	024	111111	48-56	*	73***	Black	Alcoholism
America (02)	45	111111	48-56	*	71***	White	Diabetes
India (01)	023	333333	48-56	*	72***	White	Flu

CFD ϕ_1 assures that for any person in the India (country code 01) with Area Code (AC) 023, the disease of the customer must be Diabetes, as forced by its pattern tuple (01, 023||Diabetes) correspondingly for ϕ_2 and ϕ_3. These cannot be expressed as FFDs. CFDs are recently developed for efficient privacy preserving and best semantic relationship among the conditions. They extend normal Functional Dependencies (FDs) by the implementation of semantically related constants to every Conditional Functional Dependency (CFD). CFD-based cleaning methods to be effective in practice, it is necessary to have techniques in place that can automatically discover or learn CFDs from micro table data instance D.

Description of CFD: Consider a sample relation ‘r’ for microdata table with instance ‘D’ of a schema Relation ‘R’, an algorithm for CFD aims to find CFD defined over R that hold on r. CFD functions don’t return best privacy result for all instances D to hold relation R because it contains redundant and irrelevant semantic and is unreasonably large. Thus, the desire to discover a non-redundant set of CFDs simply, from which all CFDs based on the relation R over r via implication examination. Moreover, it removes irrelevant patterns or CFD from original data instance D. A CFD $\varphi = (X \rightarrow Y, t_p)$ over R is assumed to be insignificant if $Y \in X$. If φ is trivial, either it is satisfied with every instance over the relation R (e.g., when $t_p[Y_L] = t_p[Y_R]$)) or else, it is satisfied by none of the instances, where, tuple ‘t’ such that $t[X] \leq t_p[X]$ (e.g., if $t_p[Y_L]$ and $t_p[Y_R]$)) are different constants).

A constant CFD ($\varphi = (X \rightarrow Y, t_p \| a)$) is supposed to be present left concentrated on relation r if for every $Z \subseteq X$, $r \nvDash (Z \rightarrow A, (t_p(Z) \| a))$. A variable CFD $(X \rightarrow Y (t_p \| \_)))$ is left-reduced on r if and;

- The $r \nvDash (Z \rightarrow A(t_p(Z) \| \_))$ for every appropriate subset $Z \subseteq X$
- The $r \nvDash (X \rightarrow A (t'_p(Z))))$ for every t'_p with $t_p << t'_p$. Instinctively, it assures the following conditions to satisfy CFD
- Any one of the Left Hand Side (LHS) attributes can be removed
- Any one of the constants in its Left Hand Side (LHS) sample or patterns can be “improvement” to ‘_’, i.e., the pattern $t_p(X)$ is the minimality of patterns

A minimal CFD $\varphi = (X \rightarrow Y, t_p)$ over R is assumed to be a nontrivial, left-reduced CFD such that $r \nvDash$”. Instinctively, a smallest CFD is non-redundant. It is not appropriate to each transaction; it becomes complicated to assure the privacy, consequently proposed a frequent CFD.

Frequent Conditional Functional Dependency (CFD): The support frequent pattern of a CFD $\varphi = (X \rightarrow Y, t_p)$ in relation, specified by (φ, r) is defined to be the set of tuples t in relation r from micro table instance D such that $t[X] \leq t_p[X]$ and $t[Y] \leq t_p[Y]$, i.e., tuples corresponding to the pattern of φ. For a normal number $k \geq 1$, a CFD is supposed to be k-frequent in patterns in the relation r for instance D if $sup(\varphi, r) \geq k$. For instance, two different Conditional Functional Dependencies (CFD) ϕ_1 and ϕ_2 of Example 1 are 3 and 2-frequent, correspondingly. Alternatively, a k-frequent CFD φ in r is a CFD that should hold on r, i.e., $r \nvDash \varphi$ and furthermore, there should be adequately required k observe tuples in r relation that equivalent the pattern tuple of φ.

Literature review: Privacy preserving data publishing data (Fung *et al*., 2010) with FFD has been initialized by Wang and Liu (2011). It recognized the privacy attack that makes use of functional dependencies as the opposition information and overcome the problems of privacy attack. In this research, only numerical data are considered, it can be extended to support both categorical data and numerical data through grouping approaches. Earlier research simply makes use of intersection-grouping as its grouping approach, which results in more information loss. Consequently, two new grouping strategies are introduced that can considerably result in lesser information loss ratio than the earlier anonymization techniques and evaluate the result of each grouping strategy information loss.

Wang and Wang (2013) showed that the result of k-anonymous table to protect privacy, however still it has several privacy problems due to lack of assortment in Sensitive Attributes (SA). In particular, the privacy protection result doesn’t rely on the size of the ECs of QI attributes which enclose tuples to be the same on individual attributes. As an alternative, it is resolved through the distribution of different SA values associated with every EC. In order to overcome the problems of k-anonymity, the idea of l-diversity was introduced. The results l-diversity method guarantees better privacy protection than clustered k-anonymity (Loukides and Shao, 2008).

In order to address the some problem of the anonymization, efficient bottom up partition with information technique was proposed by Hoang *et al*. (2012) with numerical attributes. Unlike the indiscriminate genetic progress and the bottom-up generalization, the present research work constructs a progressive simplified procedure which the user can step throughout to verify a preferred trade-off privacy and accuracy. Substantial research that considers the adversary information together with data correlations is observed by Kifer (2009). In particular, both works are the primary correlation based tuples to measure adversary information.

From the result, if the privacy correlation values exist, then privacy leakage occurs on the published dataset and it degrades the result of privacy utility. The researcher demonstrated the result of attacker based on privacy correlation values of the sanitized data set. Though, these works identify the privacy attacks by data correlation measures but none of the above mentioned works provided a result to protect the microdata for the attacks generated for both CFD and FFD. In order to protect privacy against attacks generated for CFD and FFD, similar patterns are identified from the microdata table. Pattern mining methods are efficient in mining the CFD and FFD based patterns for microdata table.

MATERIALS AND METHODS

Conditional functional dependencies with improved FP growth and construction of the initial partitions for the bottom-up approach with skew normal frequency distribution: Publishing microdata with the protection of privacy have become more focused in recent years. Simply, removing the quasi identifiers or changing the QI values are inadequate to defend privacy (Rastogi *et al.*, 2007). In this study, the preserved privacy model is considered to publish data. It can successfully protect privacy against the attacks generated for FFD and CFD. In this (d, l) Inference Models cannot protect privacy attacks against Conditional Functional Dependencies (CFDs), therefore proposed system move towards the direction of privacy-preserving publishing microdata model that includes CFDs. In order to successfully find the patterns, besides CFD, improved FP-growth technique is proposed which extracts frequent patterns which does not include candidate generation for CFD and necessitate two scan results for data instance D for relation r to attain an extremely compact frequency-descending tree structure (FP-tree). The proposed work presents a bottom-up approach driven by the skew normal frequency distribution. It is based on the shape mixture of skew-normal with Sensitive Attribute (SA) and gives an interesting point of view of the (d, l) Inference Model for partition of the data patterns. The skew-normal likelihood function does not converse any tools for posterior calculation. Each Sensitive Attributes (SA) partition makes the dataset into smaller groups to minimize high information loss. In order to perform the (d, l) Inference Model, the CFD should be thoroughly understood and how it differs from FFD, then formally, CFD mining pattern algorithm is defined for given micro table and then the (d, l) Inference Model is proceeded that combats the CFD. Then, the partition methods are performed. In earlier researches, FP-growth technique which mines the CFD patterns without generation of candidate patterns for CFD requires two scanning processes to achieve higher pattern mining results. During the first scan, it develops a listing of all frequent patterns from CFD in descending order based on the number of occurrence level. According to the frequency based occurrence result, the second scan produces results of CFD patterns from a table. But this method does not support dynamic dataset transaction, so an improved frequent pattern mining algorithm is presented for dynamic dataset.

Compact Frequent Pattern Growth Branch Sort algorithm (CFPGBS) for pattern mining: The improved algorithm follows the process of compact pattern tree (CP-tree), on the different, construct like same as a FP tree for CFD like compressed frequency-descending tree construction with a single-pass of a transaction database only in the microdata table (D). Initially, transactions in Table 3 are inserted into the CP-tree according to a predefined item order one by one with equivalent CDF $\varphi = (X \rightarrow Y, t_p)$ for each and every transaction. Each and every SA, QI attributes order of a CP-tree is maintained by a list called an I-list.

After inserting some of the transactions with CFD, if the attributes order of the I-list deviates from the present frequency-descending attributes order to a particular level, the CP-tree is dynamically simplified by the present frequency-descending characteristic order in the transaction table and the I-list modernize the attributes sort with the present list.

- Insertion phase: initially, it scans the database instance D, inserts attributes into the tree of order I-list and modernize the occurrence count of particular attributes in the I-list
- Restructuring phase: in this phase, the order list is reorganized for each attributes in the I-list according to frequency-descending sort of attributes and restructure the original tree structure based on new I-list order

These two phases are dynamically carried out in alternating manner for every pattern in the CFD. Then, the proposed tree restructuring technique is described and its performance characteristics are evaluated. Initially, the CFD pattern mining tree starts with null nodes. Like FP-tree, CP-tree contains nodes on behalf of a pattern set and nodes with the highest support of CFD pattern is considered as the root node. It follows the FP-tree structure method and sorting is performed to insert CFD patterns into a tree structure. Tree reformation is performed through Branch Sorting Method (BSM) and primary method obtains the I_{sort} by rearranging the attributes in I that corresponds to every microdata table transaction.

It is similar to an array based method that performs the branch-by-branch reformation procedure from the root node T of the tree. The tree generates the pattern for micro table and mines the best CFD patterns to achieve (d, l) Interference Model, the each sub-pattern tree under every child of node for patterns is considered as a branch. Therefore, a tree T contains as many patterns as the number of branches and each attributes in the patterns is the number of children it has under the root node. Each branch may consist of several paths to CFD. While restructuring a branch node, every pattern in the branch should follow the new sort array which eliminates the previous sorting array.

The new sort array would be stored in the temporary storage. However, while processing a path for CFD, if the best path is determined in the current process, the selected best pattern path is taken out from table and no sorting process is carried out. Finally, the restructuring mechanism is completed when every attributes (nodes) are processed which produces the final T_{sort}. Merge Sort Method is proceeded to sort both attributes in the item I and any unorderd CFD pattern path in T. The tree restructuring CFD mechanism using BSM Method is discussed in the study. Let P1 = $\{\alpha_1, \alpha_2,..., \alpha_n\}$ be a pattern path in T, where α_1 and α_n are the immediate next node (attributes) to the root and the leaf node (attributes) of the path, correspondingly.

P_1 is defined as a sorted pattern path that is best pattern for CFD if all items represented by α_i where I∈[1, n] and a = {Country(CO), Area Code (AC), Hospital (H), Age (A), Gender (G), Zipcode (ZC), Race (R) and Disease (D)} are found in I_{sort} order. Let a be attributes in the tree T and α_{clist} refers to its children list for CFD patterns. Let size α_{clist} be the size of this children list for every CFD pattern in the tree T. The node is considered as branching node for patterns (CFD) if the size of the nodes is greater than one $\alpha_{clist}>1$.

Based on the above definitions, the following imperative corollary is determined for usage of sorted pattern path during tree restructuring. $P_1 = \{\alpha_1, \alpha_2,..., \alpha_n\}$ which is a path in T, in which α_1 and α_n are the immediate next node to the root and the leaf node of the path, respectively. Let α_k be a branching node for each CFD patterns and $P_2 = \{\alpha_1, \alpha_2,..., \alpha_l, \alpha_{l+1},..., \alpha_m\}$ be another pattern path division with similar prefix $\{\alpha_1, \alpha_2,..., \alpha_k\}$ as in P_1 where k = L if P_1 is a sorted pattern path then sub-path $\{\alpha_1, \alpha_2,..., \alpha_l\}$ in P_2 is also sorted only if no attributes between $\alpha_{l+1},..., \alpha_m$ possesses an attribute rank less than the of α_i.

Based on the above definitions and corollary, the CFPGBS algorithm is shown below. The I_{sort} ascending order for every attributes is constructed from the Table 4. Then all the nodes of T for transcation with CFD are checked one after another (line 2) for any unsorted path (line 4) and, if found, the CFD pattern pathway is arranged in sorting order (line 6) and then the patterns are inserted into T (lines 19-22). Consequently, when there is no division left to be progressed, tree T is entirely restructured and the newly restructured tree is the sorted list CDF for every pattern (line 7).

During tree restructuring, if any path is found to be a sorted CFD pattern path (line 4), it does not only leave out the sort process for that CFD pattern path but furthermore progress this status information about the pattern path to everyone branching nodes in the same pattern path of the entire branch. Consequently, the remaining pattern paths in the similar branch are also sorted. The attributes of the sub-pattern paths from the leaf node pattern and the division node attributes for each sorted pattern path are checked to determine whether any attribute in the sub-pattern path contains a rank result less than or equal to the CFD patterns (line 11).

If no such attributes are found, the sub-pattern path from the root to the branching node is maintained as such which is considered as common prefix CFD pattern path, as it simply process the CFD sub pattern path from the division node to the leaf nodes with CFD patterns (lines 12, 13, 14, 16). Or else, the entire pattern CFD path from the leaf pattern path nodes to the root patterns path (line 15) is sorted (line 16). When every sub-pattern paths (line 10) from the branching node are adjusted, it is moved to the next accessible branching node (line 9).

Table 4: Transaction table with CFD of hospital patient discharge data

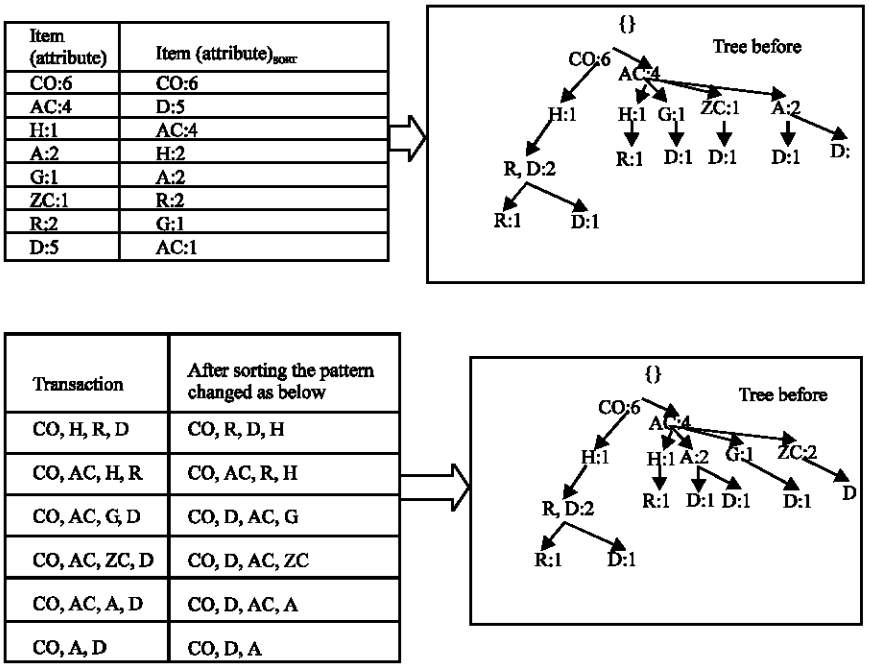
Transaction (conditions)	CFD	Patterns
CO,H,R,D (1)	CO, H→R, Disease	{01|02,111111,_||_}
CO,AC,H,R (2)	CO, AC, H→R	{01,023, -, White}
CO,AC,G,D (3)	CO, AC, G→Disease	{01,_,_, Alcoholism||flu}
CO,AC,ZC,D (4)	CO, AC, ZC→Disease	{02,45,_, Diabetes}
CO,AC,A,D (5)	CO, AC, A→Disease	{01,023,39||54, Diabetes||flu}
CO,A,D (6)	CO, A→Disease	{01||02,37-45||48-56, White}

```
Input: Original Tree (T) with CDF, Items (attributes)
Output: $T_{sort}$ and $I_{sort}$
Compute $I_{sort}$ from item in frequency-descending order using merge sort method
For each branch $B_i$ in tree (T)
For each unprocessesd path pattern $P_i$ in $B_i$
If $P_j$ js a sorted path
Process_Branch ($P_j$)
Else sort_path ($P_j$)
Terminate when all the branches are sorted and output $T_{sort}$ and $I_{sort}$
Process _Branch ($P_i$)
{ For each branching node $n_b$ in P from leaf node
For each sub_path from $n_b$ to $leaf_k$ with $k \neq P$
If item (attribute) rank of the nodes between from c are greater than that of $n_b$
P = sub path from $n_b$ ti $leaf_k$
If P is sorted path
Process _Branch(p)
Else P = path from root node to $leaf_k$
Sorted_path(P) }
Sorted _path(Q)
Reduce the count of all path nodes of Q by the values of $leaf_Q$ count
Using the merge sort in Q array to $I_{sort}$ order
Delete all nodes from Q count has zero value
Insert Sorted Q into T at the location from the location where it was taken
```

In the above steps followed for each and every CFD is described as follows with an example. An item set pair (X→Y, t_p) is defined in which $X \subseteq attr(R)$ and t_p is a CFD pattern over X. A CFD $\varphi = (X \rightarrow Y, t_p)$ over R is assumed to be insignificant if $Y \in X$. If φ is trivial, either it is satisfied by every one of instance over the relation R (e.g., when $t_p[Y_L] = t_p[Y_R]$) or else it is satisfied by none of the instances in which a tuple t is present such that $t[X] \le t_p[X]$ (e.g., if $t_p[Y_L]$ and $t_p[Y_R]$) are different constants). The support value for each CFD patterns in the relation r is specified by supp X→Y, t_p, r.

The transaction table and the equivalent patterns are defined to mine the efficient patterns to perform (d, l) Interference Model relating generators to their closures l-diversity and then partition the dataset into groups with LSKNAPD. This can be done by collecting the non closed CFD patterns subset for given closed CFD patterns that have the same support value. Consider the following transaction table with CFD of hospital patient discharge data to perform the pattern mining for each and every CFD, finally the best CFD are mined using Compact frequent pattern growth branch sort algorithm (CFPGBS).

Figure 1 and 2 shows how is the transaction is altered before and after CFD in the database. The sorting technique sorts the first path {CO:1, H:1, R:1} of the branch which is an unsorted path, it is removed from the tree and then sorted using the merge sort technique into a temporary array to the order {CO:1, R:1, D:1, H:1} by satisfying I_{sort} order and then again inserted into tree T in sorted order. Figure 1b shows the result of pattern mining tree structure for CFD after sorting the first path as well as all the pattern paths in CFD. The tree maintains complete database information instance D in a more compact manner.



Item (attribute)	Item (attribute)$_{SORT}$
CO:6	CO:6
AC:4	D:5
H:1	AC:4
A:2	H:2
G:1	A:2
ZC:1	R:2
R;2	G:1
D:5	AC:1

Transaction	After sorting the pattern changed as below
CO, H, R, D	CO, R, D, H
CO, AC, H, R	CO, AC, R, H
CO, AC, G, D	CO, D, AC, G
CO, AC, ZC, D	CO, D, AC, ZC
CO, AC, A, D	CO, D, AC, A
CO, A, D	CO, D, A

Fig. 1: Tree after and before sorting with transaction table

From this sorting result, it is observed that the patterns CO, AC, G, D are not required as it becomes lesser privacy when compared to other tuples in the result. So, it is removed from the table and then the diversity checking is also changed as shown in Table 5. It becomes a 2-diversity and the area code also has to be changed because when user knows the area code he/she can easily identifies the patient data. Now diversity and CDF functions are also applied to Table 5, then conditional dependency functions is chosen from transaction table after sorting it which satisfies the condition 1, 2, 4, 5 and 6. Now the table is changed as shown in Table 6.

Log-skew-normal alpha-power distribution: In this study, a new frequency distribution model is presented to group the partition data from the result of (d, l) Inference Model with CFD. The number of occurrences in CFD that satisfies the d-closness, for example, it is defined as two, then totals two number of groups with (d, l) Inference Model is used. It is the new set of distributions for conditional dependency pattern based on a Log-Skew-Normal Alpha-Power distribution (LSKNAPD) with skew functions; it performs in efficient manner when compared to other distribution function such as Log-Normal (LN) model and Log-Skew-Normal (LSN) Model.

In order to analyze the result of privacy protection d-closeness it is necessary to deal with the condition of close frequency for every attributes. In particular, it is observed that s_1 and s_2 two data attribute values that are close to distribution uniform and skew distribution d-close if $|C_1\text{-}C_2|\leq d$ where C_1 and C_2 are the count frequency of condtions that related to s_1 and s_2. Consider a group G that consists of a set of distinct l sensitive data values which are distribution uniform and skew distribution d-close, if for any two values S_i, $S_j\in G$, where s_i, s_j are d-close. The detailed definition of distribution (d) d-closeness is described. Then, the partitions is defined as follows. Given a set of distinct values $S = \{S_1,..., S_n\}$ a partition scheme of S is a set of segments $\{p_1,..., p_t\}$.

The positive random number of the group G in the R^+ with quantitative log skew-normal alpha-power distribution function analysis for every group through parameters λ and α, it is transformed to log transformation function Z = log(G). This is denoted by G~LPSN(λ, α). The Probability Density Function (PDF) of any random variable in the group G with distribution LPSN(λ, α) is given by:

$$\varphi_{LPSN}(g,\lambda,\alpha)=\frac{\alpha}{g}\phi_{SN}(\log(g)):\lambda)\{\phi_{SN}(\log(g);\lambda)\}^{\alpha-1},\quad g,\alpha\in R^{+}\text{ and }\lambda\in R \tag{3}$$

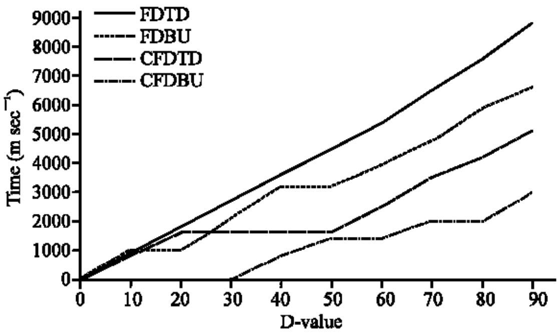


Fig. 2: Time performance comparison U distribution (FDTD, FDBU, CFDTD, CFDBU)

Table 5: Transaction table after removal of gender from the hospital patient discharge data

Quasi identifier(QI)					Sensitive Attributes (SA)	
CO	AC	H	A	ZC	R	ICD-9-CM
India (01)	022	111111	39	71000	Asian	HIV
India (01)	023	111111	41	72001	White	Diabetes
America (02)	45	222222	48	73021	Black	Diabetes
India (01)	024	222222	43	72001	Black	Flu
India (01)	024	111111	48	73020	Black	Alcoholism
America (02)	45	111111	51	71000	White	Diabetes
India (01)	023	333333	56	72000	White	Flu

Table 6: Transaction table after CFD with removal of gender from the hospital patient discharge data

Quasi identifier(QI)					Sensitive attributes (SA)	
CO	AC	H	A	ZC	R	ICD-9-CM
India (01)	*	111111	≤50	7****	Asian	HIV
India (01)	*	111111	≤50	7****	White	Diabetes
India (01)	*	222222	≤50	7****	Black	Flu
India (01)	*	111111	≤50	7****	White	Diabetes
America (02)	*	222222	>50	7****	Black	Diabetes
America (02)	*	111111	>50	7 ****	White	Diabetes
India(01)	*	333333	>50	7****	White	Flu

The Cumulative Distribution Function (CDF) of the LPSN Model is specified by:

$$F_G(g,\lambda,\alpha)=\{\phi_{SN}(\log(g);\lambda)\}^{\alpha},g,\alpha\in R^{+} \tag{4}$$

According to Eq. 4, the inversion method can be used to generate a random variable with frequency distribution LPSN(λ, α). Thus, if U~(0, 1) is a uniform random variable in the group G = exp {ϕ_{ISN}($U^{-1/\alpha}$, λ)} has LPSN distribution of the parameters λ and α where ϕ_{ISN} denotes the inverse function result of SN frequency distribution in every attributes in the group. When λ = 0 and α = 1, the LPSN distribution is equivalent to the LSN distribution [φ_{LPSN}(g, λ, α) = φ_{LSN}(g, 0, 1, α)] the LPSN distribution is the similar to LN distribution. It is more efficient than other two distributions LN and LSN. The rth moment of the random variable in the group G with LPSN distribution can be specified as:

$$\mu_r=E(G^{r})=\alpha\int_0^1\{\exp[r\varphi_{ISN}(g,\lambda)]\}g^{\alpha-1}\,dg \tag{5}$$

Let:

$$\mu_r^{'}=E(G-E(G))^{r},r=2,3,4,$$
$$\mu_2^{'}=\mu_2-\mu_1^2,\mu_3^{'}=\mu_3-3\mu_2\mu_1+2\mu_1^3,$$
$$\mu_4^{'}=\mu_4-4\mu_3\mu_1+6\mu_2\mu_1^2-3\mu_1^4$$

The variance, coefficient of variation, skewness and kurtosis are given by:

$$\sigma_G^2=Var(G)=\mu_2^{'},CV=\frac{\sqrt{\sigma_G^2}}{\mu_1},\sqrt{\beta_1}=\frac{\mu_3^{'}}{|\mu_2^{'}|^{3/2}}\text{ and }\beta_2=\frac{\mu_4^{'}}{|\mu_2^{'}|^{2}}$$

Let PSN (ξ, η, λ, α) indicate the transformation of power distribution for Group G PSN (λ, α) where $\xi \in R$, $\eta \in R^{+}$ and G = $\xi + \eta$Z. If X has a distribution of parameters in the group with power frequency distribution occurs for QI and SA PSN (ξ, η, λ, α) then the extension of frequency distribution for log skew normal transformation is follows X = log(G) where $\xi \in R$, $\eta \in R^{+}$. Then the density of the group G is given by:

$$\varphi_{LPSN}(g;\xi,\eta,\lambda,\alpha)=\alpha\varphi_{LPSN}(g,\xi,\eta,\lambda)\left\{\phi_{SN}\left(\frac{\log(g)-\xi}{\eta}\right)\right\}^{\alpha-1} \tag{6}$$

The special case of frequency distribution when λ = 0, obtaining the density,

$$\varphi_{LPSN}(g;\xi,\eta,\lambda,\alpha)=\frac{\alpha}{\eta g}\varphi_{LPSN}\left(\frac{\log(g)-\xi}{\eta}\right)\left\{\phi\left(\frac{\log(g)-\xi}{\eta}\right)\right\}^{\alpha-1} \tag{7}$$

This density is denoted as~LPSN$_{\lambda=0}$(ξ, η, α). The LSN Model follows the general procedure of the LN Model so it is called as generalized LN model. The expansion result of both LSN and LN Models includes the following steps: λ = 0 and α = 1 in the LPSN model to LN models, LN model doesn't generate MGF for frequency distribution. Since MGF satisfies the property, $M_{\alpha G+b}(t)$ = exp(bt) $M_G(\alpha t)$.

Then, it is an adequate condition to every standard case of the group in frequency distribution function. For any fixed values $\alpha = \alpha_0 > 0$ and $\lambda = \lambda_0$ the MGF of G can be specified as:

$$\begin{aligned}M_G(t)&=E\left(e^{tg}\right)\\&=\int_0^{\infty}e^{tg}\varphi_{LPSN}(g;\lambda_0,\alpha_0)dg\\&=\int_0^{\infty}\frac{\alpha_0}{g}e^{tg}\varphi_{SN}(\log(g);\lambda_0)\Phi_{SN}(\log(g);\lambda_0)\}^{\alpha_{0-1}}dg\\&=\int_0^{\infty}\frac{\alpha_0}{g}e^{tg}\varphi_{SN}(\log(g);\lambda_0)\Phi_{SN}(\log(g);\lambda_0)\Big\}^{\alpha_0-}dg\\&=\int_0^{\infty}h(g,t,\lambda_0,\alpha_0)g(g,t,\lambda_0,\alpha_0)dg\end{aligned}$$

With:

$$h(g,t,\lambda_0,\alpha_0)=(2\alpha_0/g)\,e^{gt}\,\phi\Big(\log(g)\big\{\Phi(\lambda_0\log(g)\big\}>0,$$
$$g(g,\lambda_0,\alpha_0)=\phi_{SN}(\log(g)\{\lambda_0\}>0\text{ to all }g>0$$

when t>0 is fixed prove that:

$$J_{(\lambda_0,\alpha_0)}=\int_0^{\infty}h(g,t,\lambda_0,\alpha_0)g(g,\lambda_0,\alpha_0)$$
$$dg=\infty\text{ For all }\lambda_0\in R\text{ and }\alpha_0\in R^{+}$$

So, when g→∞ have the asymptotic estimate:

$$\log\big(\Phi(\lambda_0\log(g))\big)\approx\frac{1}{2}(\lambda_0\log(g))^2$$

Then, it is assumed that the log function of group $\log(\alpha_0)<\infty$ where g→∞ should be:

$$\log\big(h(g,t,\lambda_0,\alpha_0)\big)-\log(\alpha_0)\approx\frac{1}{2}\log\left(\frac{2}{\pi}\right)-$$
$$\log(g)+tg-\frac{1}{2}(\lambda_0^2+1)(\log(g))^2\to\infty$$

Now because g(g, λ_0, α_0)→1 when group tends to infinity g→∞ then says that $J_{(\lambda_0,\alpha_0)}\to\infty$ when g→∞. The maximum likelihood estimation and observed result for QI and SA groups provide the expected matrix information for the parameters of the LPSN (ξ, η, λ, α) Model considered. For a random sample of size n, G_1, G_2,..., G_n with G_i~LSPN(ξ, η, λ, α), the most likelihood result for Group G, θ = (ξ, η, λ, α) , can be expressed by:

$$I(\theta,G)=n\big(\log(\alpha)-\log(\eta)\big)-\sum\nolimits_{i=1}^{n}\log(g_i)-\frac{1}{2}\sum\nolimits_{i=1}^{n}Z_i^2+\sum\nolimits_{i=1}^{n}\log\,(\Phi(\lambda Z_i)\}+(\alpha-1)\sum\nolimits_{i=1}^{n}\log\,(\Phi_{SN}(Z_i,\lambda))]$$

where, Z_i = log(g_i)-ξ/η. The elements of the score function for SA and QI are given by:

$$U(\xi)=\frac{1}{\eta}\sum_{i=1}^{n}Z_i-\frac{\lambda}{\eta}\sum_{i=1}^{n}w_i-\frac{\alpha-1}{\eta}\sum_{i=1}^{n}w_{1i}$$

$$U(\eta)=\frac{-n}{\eta}+\frac{1}{\eta}\sum_{i=1}^{n}Z_i^2-\frac{\lambda}{\eta}\sum_{i=1}^{n}z_iw_i-\frac{\alpha-1}{\eta}\sum_{i=1}^{n}w_{1i}Z_i$$

$$U(\lambda)=\sum_{i=1}^{n}Z_iw_i-\sqrt{\frac{2}{\pi}}\frac{\alpha-1}{1+\lambda^2}\sum_{i=1}^{n}w_i(\lambda)$$

$$U(\alpha)=\frac{n}{\alpha}+\sum_{i=1}^{n}\log(\Phi_{SN}(Z_i;\lambda)$$

Where:

$$w=\frac{\phi(\lambda z)}{\phi(\lambda z)},w_1=\frac{\phi_{SN}(Z)}{\Phi_{SN}(Z;\lambda)}\text{ and }w(\lambda)=\frac{\phi\left(\sqrt{1+\lambda^2}z\right)}{\Phi_{SN}(z;\lambda)}$$

The score result for each QI and SA is obtained by equating these partial derivatives of attributes to zero. The Maximum Likelihood Estimators (MLEs) are the efficient methods to solve the result of score result for each QI and SA. It is usually achieved throughout the iterative numerical methods. The LSKNAPD derives the greatest frequency distribution for every group G that consists of a set of l-distinct data values that is close to d-close if for each two sensitive values S_i, S_j∈G, where S_i, S_j are d-close. After finding the frequency count for each and every CDF that are closer to d-closeness, bottom up partition approach is applied to combine the same frequency values that are closer to every other as one group through (d, l) Inference Model.

Bottom-up approach: In bottom up approach, all the quasi identifier results are partitioned into only one result; the working principle of the bottom up approach is opposite to top down approach. Primarily, the set of all individual Sensitive Attribute (SA) values S = {s_1,..., s_n} is partitioned into set of many disjoint segments based on the result from Log-skew-normal alpha-power distribution, for l distinct values, except for the last one that might contain more than one l distinct senstive attribute values. For each partition S {s_1,..., s_n}, the number of removed tuples is determined by applying Log-skew-normal alpha-power distribution on partition P.

Then starting from the initial division, every two adjacent ones are integrated from partition result LSKNAPD P'_i = P_i∪P_{i+1} and the results are compared with frequency LSKNAPD result of information loss IL(P'_i) with (P_i) + IL(P_{i+1}). If IL(P'_i)≥IL(P_i) + IL(P_{i+1}), it will not be combined. Otherwise (P_i) and (P_{i+1}) will be integrated to construct P'_i. The above process is repeated until no P_i can be combined. Bottom-up approach integrates different groups and is concerned with the frequency distribution based grouping, which always constructs (d, l) inference groups. So, entire partitions result should satisfy (d, l) Inference Model with CFDS.

The result of initial Log-skew-normal alpha-power distribution partition simply considers the number of different sensitive values, but it does not consider the number of tuples; it becomes difficult to partition the larger dataset with frequent sensitive attribute values and high information loss. In order to reduce such information loss, each partition is split into smaller groups by referring grouping methods (Wang and Liu, 2011). In particular, the partition result from distribution function k is the number of different l-distinct data values. It is basic that k≥l, the condition parameter of the (d, l) Inference Model. Then the tuples in the partition P results are bucketized by using hashing values.

It requires alteast k different values for k hashed buckets; therefore, those different values will not be hashed into the similar bucket. After completeion of bucketization step in which group QI chooses k tuples from k buckets, one from every one. These QI grouping procedures are repeated until there is only one unselected tuple present in bucket.

Then, the remaining tuples of every bucket are grouped into one QI-group. Let f_{min} be the minimum frequency assessment in partition P. At the end of the stage, it contains f_{min} constructed QI-groups, out of them f_{min}-1 includes of k(k≥l) distinct values of single occurrence with the value S_i of f_i-f_{min} + 1 frequency, where f_i is the frequency of sensitive attribute S_i. The result should satisfy (d, l) Inference Model with CFD.

RESULTS AND DISCUSSION

In this study, a widespread set of experiments is carried out to assess both the usefulness and efficiency of Conditional Functional Dependency (CFD) based (d, l) Inference Model. Experimental results are evaluated for l, d parameters, uniform and skew data distribution with number of CFD and FFD on both the accuracy and effectiveness of the generalized data. The effectiveness of the two partition approaches are compared with the existing partition approach. CENSUS dataset contains the personal information of real dataset. Census dataset that having size of 100 K. Census dataset totally consists of 15 attributes and attributes information are mentioned in Table 7, records samples of each amercian dataset samples also mentioned in Table 8.

In this research consider 1500 samples as input for privacy preserving (d, l) Inference Model. The attributes, mentioned in Table 1 majorly there are five attributes such as age, education, marital status, education-num and work class be considered as Sensitive Attributes (SA) and remaining 10 attributes such as fnlwgt, occupation, relationship, race, sex, capital gain, capital loss, hours-per-week, native country, salary are considered as Quasi Identifiers (QI). The dataset comprises of totally six attributes where the distribution result on sensitive values is close to uniform distribution and another distribution whose sensitive values are closes to skewed distribution. Then, these two distribution data are called as the U-dis and S-dis dataset.

It is observed that the proposed bottom up approach with frequency are much faster than existing top down and bottom-up approach. It pays attention for privacy protection when FFD and CFD are present. It simultaneously, increases the distribution values D for each and every partition method for both FFD and CFD. The time performance compares U and S distribution (FDTD: Functional dependency top-down, FDBU: Functional dependency bottom-up, CFDTD: Conditional Functional dependency Top down, CFDBU: Conditional Functional dependency Bottom up) as shown in Fig. 2 and 3, respectively. Figure 4 shows the time performance comparison of the grouping methods IG, CG and DG. Figure 4 presents the results of time comparison (FDTD, FDBU, CFDTD, CFDBU).

Table 7: American Adult dataset attributes information

Attribute name	Attribute informations
Age (A)	Continuous
Workclass (WC)	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Fnlwgt (FW)	Continuous
Education (ED)	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5-6th, Preschool
Education-num (EDN)	Continuous
Marital-status (MS)	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation (O)	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Relationship (RL)	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Race (R)	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
Sex (S)	Female, Male
Capital-gain (CG)	Continuous
Capital-loss (CL)	Continuous
Hours-per-week (HPW)	Continuous
Native-country (NC)	United-States etc
Salary (S)	≤50000 and ≥50000 thousand

Table 8: Adult dataset samples

Values	Adult datasets
39	State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, ≤50K
50	Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, ≤50K
38	Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, ≤50K
53	Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, ≤50K
28	Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, ≤50K
37	Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, ≤50K
49	Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, ≤50K
52	Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 45, United-States, >50K
31	Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White, Female, 14084, 0, 50, United-States, >50K
42	Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 5178, 0, 40, United-States, >50K
37	Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Black, Male, 0, 0, 80, United-States, >50K
30	State-gov, 141297, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Islander, Male, 0, 0, 40, India, >50K
23	Private, 122272, Bachelors, 13, Never-married, Adm-clerical, Own-child, White, Female, 0, 0, 30, United-States, ≤50K
32	Private, 205019, Assoc-acdm, 12, Never-married, Sales, Not-in-family, Black, Male, 0, 0, 50, United-States, ≤50K
40	Private, 121772, Assoc-voc, 11, Married-civ-spouse, Craft-repair, Husband, Asian-Pac-Islander, Male, 0, 0, 40, ?, >50K
34	Private, 245487, 7th-8th, 4, Married-civ-spouse, Transport-moving, Husband, Amer-Indian-Eskimo, Male, 0, 0, 45, Mexico, ≤50K
25	Self-emp-not-inc, 176756, HS-grad, 9, Never-married, Farming-fishing, Own-child, White, Male, 0, 0, 35, United-States, ≤50K
32	Private, 186824, HS-grad, 9, Never-married, Machine-op-inspct, Unmarried, White, Male, 0, 0, 40, United-States, ≤50K
38	Private, 28887, 11th, 7, Married-civ-spouse, Sales, Husband, White, Male, 0, 0, 50, United-States, ≤50K

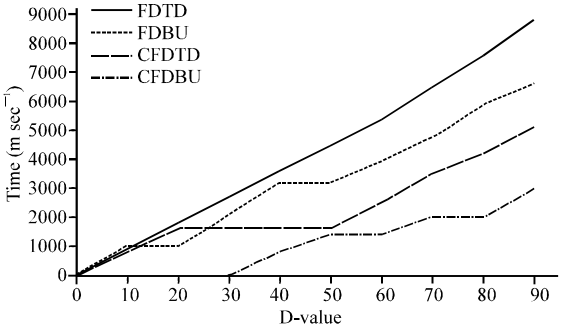


Fig. 3: Time performance comparison S distribution (FDTD, FDBU, CFDTD, CFDBU)

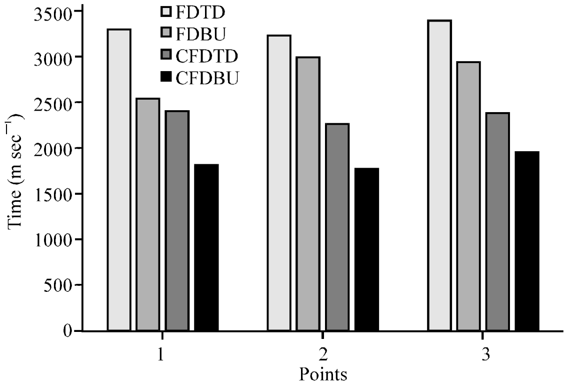


Fig. 4: Time performance comparison, CENSUS dataset for CFD, FFD

It shows that the time performance of these approaches is not associated to the frequency distribution of the data to be grouped; CG is always faster for CFDBU than FDTD, FDBU, CFDTD using FFD and CFD. The proposed LPSN Model uses the CFD with bottom up approaches. The discermibility metric is used by Jin *et al*. (2010) and has been used as the quantity in a few previous researches. In this metric, every generalized tuple is allocated a penalty of term with size of group |G| while a concealed tuple is allocated a penalty of |D| the size of the contribution census dataset. The total Information Loss (IL) for every partition method is estimated using the following formula:

$$\mathrm{IL} = \sum_{i=1}^{t} | G_i |^2 + b^* | D |$$

where, t the number of QI-groups in the result is for CFD and FFD and b is the number of removed tuples for every grouping.

The information loss of the proposed bottom-up approaches is compared with LPNA for CFDs and the previous top down and bottom approaches for FFD as shown in Fig. 5. It is observed that the bottom-up approaches with LPNA for CFDs always returns less information loss when compared with other two approaches. This proves the trade-off between privacy and effectiveness, better privacy security can be attained by compromising utility. The varied l value of the attributes for CFD and FFD with partition measures for individual distribution results evaluates the information loss.

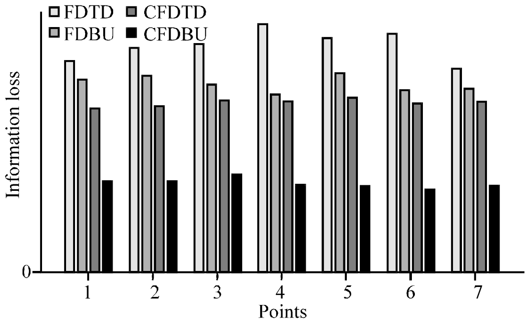


Fig. 5: Information loss comparison for census dataset

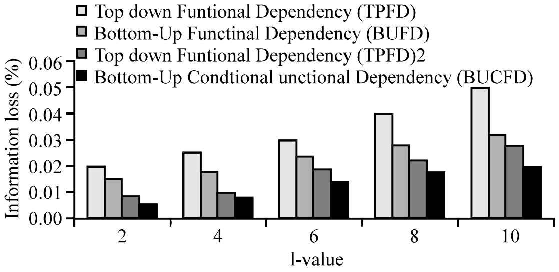


Fig. 6: Information loss comparison for U-dis dataset

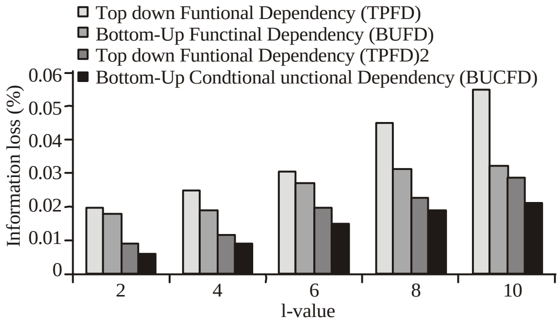


Fig. 7: Information loss comparison for S-dis dataset

Figure 6 and 7 shows that for both CFD and FFD distribution dataset, increasing l values brings additional information loss since number of the l distinctive sensitive attributes and G_i increases. It must satisfy l values for CFD and FFD that are close to the distribution function to satisfy d-closeness; secondly, larger l values result in more universal alteration on the data which incurs poorer information loss by suppression with FFD in top-down and bottom up approaches and it provides less information loss in Conditional Functional Dependency Bottom up with best frequency distribution (CFDBU).

CONCLUSION

This study investigates the problem of privacy preserving publishing of microdata considering both CFD and FFD as adversary knowledge. A Compact Frequent Pattern Growth Branch Sort Algorithm (CFPGBS) is proposed to discover minimal CFDs from a dataset when its range is large. The algorithm formulates the privacy model, (d, l) inference to protect privacy loss of information that is caused by CFDs, FFDs. An Automatic Compact Frequent Pattern Growth Branch Sort algorithm (CFPGBS) is also proposed for mining the best CFD patterns and removing the less quality CFD patterns which is considered to be an NP-hard problem. Also a compact pattern tree is developed, that captures CFD patterns information with insertion phase and provides the better pattern mining performance for CFD patterns. The construction of the initial partitions for the bottom-up approach driven is performed by Log-Skew-Normal Alpha-Power Distribution (LSKNAPD) frequency distribution function. An extensive set of experimental results show that the proposed (d, l) Inference Model that can proficiently anonymize the microdata with a low information loss against CFD. As a future research we have planned to devise a novel method for publishing multiple subtables such that each sub-table is anonymized by an existing QI-SA publishing algorithm, while the combination of all published tables guarantees all privacy rules in (d, l) diversity methods. Substantial work can be carried out by applying Boyce Codd Normal Form (BCNF) on mined CFD patterns to identify new privacy rules and novel decomposition algorithms. Designing efficient privacy preserving algorithms to defend against multiple CFDs also remains as an open research problem to be addressed in future.

REFERENCES

Chhinkaniwala, H. and S. Garg, 2014. Privacy gain based multi-iterative k-anonymization to protect respondents privacy. Int. J. Comput., 3: 85-92.

Cormode, G., L. Golab, K. Flip, A. McGregor and D. Srivastava *et al*., 2009. Estimating the confidence of conditional functional dependencies. Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, June 29-July 02, 2009, ACM, New York, USA., ISBN: 978-1-60558-551-2, pp: 469-482.

Fung, B.C.M., K. Wang, R. Chen and P.S. Yu, 2010. Privacy-preserving data publishing: A survey of recent developments. ACM Comput. Surv., Vol. 42. 10.1145/1749603.1749605

Hoang, A.T., M.T. Tran, A.D. Duong and I. Echizen, 2012. An indexed bottom-up approach for publishing anonymized data. Proceedings of the 2012 Eighth International Conference on Computational Intelligence and Security (CIS), November 17-18, 2012, IEEE, Guangzhou, China, ISBN: 978-1-4673-4725-9, pp: 641-645.

Jin, X., M. Zhang, N. Zhang and G. Das, 2010. Versatile publishing for privacy preservation. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 25-28, 2010, ACM, New York, USA., ISBN: 978-1-4503-0055-1, pp: 353-362.

Kifer, D., 2009. Attacks on privacy and deFinetti's theorem. Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, June 29-July 02, 2009, ACM, New York, USA., ISBN: 978-1-60558-551-2, pp: 127-138.

Loukides, G. and J.H. Shao, 2008. An efficient clustering algorithm for k-anonymisation. J. Comput. Sci. Technol., 23: 188-202.

Martin, D.J., D. Kifer, A. Machanavajjhala, J. Gehrke and J.Y. Halpern, 2007. Worst-case background knowledge for privacy-preserving data publishing. Proceedings of the IEEE 23rd International Conference on Data Engineering ICDE 2007, April 15-20, 2007, IEEE, Istanbul, Turkey, ISBN: 1-4244-0802-4, pp: 126-135.

Medina, R. and L. Nourine, 2009. A Unified Hierarchy for Functional Dependencies, Conditional Functional Dependencies and Association Rules. In: Formal Concept Analysis. Ferre, S. and S. Rudolph (Eds.). Springer Berlin Heidelberg, Berlin, Germany, pp: 98-113.

Nergiz, M., M. Atzori and C. Clifton, 2007. Hiding the presence of individuals from shared databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 12-14, 2007, Beijing, China, pp: 665-676.

Rastogi, V., D. Suciu and S. Hong, 2007. The boundary between privacy and utility in data publishing. Proceedings of the 33rd International Conference on Very Large Data Bases, September 23-28, 2007, VLDB Endowment, Austria, ISBN: 978-1-59593-649-3, pp: 531-542.

Wang, H. and R. Liu, 2011. Privacy-preserving publishing microdata with full functional dependencies. Data Knowl. Eng., 70: 249-268.

Wang, P. and J. Wang, 2013. L-diversity algorithm for incremental data release. Appl. Math. Inf. Sci., 7: 2055-2060.

Wong, R.C.W., J. Li, A.W.C. Fu and K. Wang, 2006. The (α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2006, ACM, New York, USA., ISBN:1-59593-339-5, pp: 754-759.

Xiao, X., Y. Tao and N. Koudas, 2010. Transparent anonymization: Thwarting adversaries who know the algorithm. ACM. Trans. Database Syst., 35: 1-48.