Asian Journal of Information Technology 15 (2): 347-358, 2016
ISSN: 1682-3915

Boyce-Codd Normal Form (BCNF) Based Privacy-Preserving Publishing Multiple Subtables with Conditional Functional Dependencies

[1]S. Balamurugan and [2]P. Visalakshi
[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 microdata amplifies the problem arising out of individual privacy entity. This study investigates the problem of privacy preservation hazard of mined Conditional Functional Dependency (CFD) against (d, l) Inference Model using CFPGBS methods. The major problem of the above mentioned methods is that, it protects privacy only for the entire table without considering the attribute wise privacy for both CFDs and FFDs against the (d, l) Inference Model. In order to overcome these limitations, Boyce-Codd Normal Form (BCNF) method has been presented to facilitate publishing of multiple subtables and it is anonymized through (d, l) Inference Model however the integration of different published subtables also needs to guarantee privacy rules for CFDs. The construction of the initial partitions for the driven bottom-up approach is performed through Fuzzy Binomial Distribution (FBD). Experimental results show that the proposed BCNF (d, l) Inference Model can adeptly anonymize the microdata with less information loss as compared to (d, l) Inference Model with CFD. The effectiveness and privacy results of proposed BCNF (d, l) Inference Model with FBD is also significant in comparison to the existing Inference Model with Log-Skew-Normal Alpha-Power distribution (LSKNAPD) function.

Key words: Conditional Functional Dependency (CFD), Privacy preservation, Fully Functional Dependency (FFD), bottom up approach, Compact Frequent Pattern Growth Branch Sort algorithm (CFPGBS), Boyce-Codd Normal Form (BCNF), Fuzzy Binomial Distribution (FBD) Log-Skew-Normal Alpha-Power Distribution (LSKNAPD)

INTRODUCTION

Privacy-Preserving publishing of micro data has become an attractive research area in recent years in the field of data mining as it protects privacy of individual sensitive information. Microdata table comprises of information about individual records, agencies related information and other organizational data for e.g., medical data or census data and data used for investigation processes and other purposes. Generally, these data are represented as a table where each row corresponds to the individual data records and each column corresponds to the attributes such as Quasi Identifiers (QI), Explicit Identifiers (EI) and Sensitive Attributes (SA).

- Attributes that are considered as explicit identifiers which are easily identifiable as per social security number, address and name and so on
- Attributes values which shows the person's identify are known as Quasi-Identifiers (QI) and might contain details for e.g., zip-code, birth date and gender
- Attributes that are categorized as confidential information are known as Sensitive Attributes (SA) such as disease and salary

A number of anonymization techniques have been developed by various researchers (Wang *et al*., 2004; Bayardo and Agrawal, 2005; Xiao and Tao, 2006; Ali *et al*., 2015). Generalization (k anonymization (Sweeney, 2002) and bucketization (Machanavajjhala *et al*., 2007) (l-diversity) are two important approaches. Similarly other types of anonymization based on privacy protection such as k anonymity (Sweeney, 2002), l-diversity (Machanavajjhala *et al*., 2007), t-closeness (Li *et al*., 2007a) were also developed. K-anonymity does not offer adequate privacy against attribute discovery. It is engaged for overcoming by using the l-diversity but it

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

allows privacy protection for limited information only. If the value of l increases considerably, individualistic information loss becomes too high. This problem has been overcome by using closeness methods but still it fails to protect privacy information without taking into consideration the semantic relationship between attributes. These problems are overcome by using other anonymization methods such as (α, k) anonymity (c, k) safety (Martin *et al*., 2007), etc.

But, none of the above mentioned tasks handled the problem of FRull Functional Dependencies (FFDs) attacks and Conditional Functional Dependencies (CFDS) (Fan *et al*., 2008) privacy attacks effectively. If the attacker knows CFD and FFD then he/she can easily find confidential information regarding the individual, since these methods based on FFD (Wang and Liu, 2011) do not evaluate the semantic relationship between attributes, through (d, l) Inference Model, this is overcome by using CFDs. But still, there are certain limitations in both FFDs and CFDs against Inference Model. For instance, if the published CFD based mined microdata table is split into many subtables, it violates the privacy of the individual, since the existing methods satisfy privacy for only the entire table without taking in to consideration the attribute wise privacy preserving for CFDs. In order to overcome these problems that have been mentioned above in the following tasks, a Boyce-Codd Normal Form (BCNF) approach is proposed to protect privacy against multiple subtables for CFDs against (d, l) Inference Model. The main contributions of the study are:

- Initially, CFDs is defined clearly and how it varies from Functional Dependencies (FDs) and is applicable to the anonymization process
- Secondly, a mining method is defined for CFD against (d, l) Inference Model through CFPGBS algorithm
- Thirdly, a Boyce-Codd Normal Form (BCNF) approach is used to identify the violated privacy rule from mined CFD patterns and then decomposes the published data table into two subtables to remove the privacy violation problem
- Fourthly, a well-organized (d, l) Inference Model is presented against CFD attacks with reduced information loss. It consists of two major steps namely, the initial partition and QI-group construction. The initial partition of bottom up approach is carried out using the frequency distribution method such as Fuzzy Binomial Distribution (FBD). Then, grouping strategies are deployed as described by Hui and Liu.
- Finally, the privacy results of proposed and existing approach are evaluated through a set of experiments using census dataset. The experimental results of the census dataset shows that the proposed privacy preservation published data that corresponds to multiple subtables preservation has anonymized data efficiently with lesser information loss for CFDS than the FFDs that are available

Literature rewiew: A number of metrics have been used in studies that have been conducted previously that assess the results of privacy in published datasets. Perfect privacy is one of the fundamental metrics used in studies carried out previously (Miklau and Sucin, 2004) that assess the results of privacy in published micro data table and additionally it possibly releases any sensitive information of patient details. Large number of subsequent researchess has been also used in earlier years to improve the perfect privacy results by deploying the conjunctive query (Machanavajjhala and Gehrke, 2006). Several numbers of privacy preserving methods such as k anonymity, l-diversity and t-closeness) methods have been proposed to address the problem of different privacy preserving requirements. But, none of the above mentioned really protects privacy based on data correlations.

In order to resolve these problems and to evaluate the correlation among data tuples for published micro data table, Martin *et al*. (2007) and Rastogi *et al*. (2007) initially have considered adversary information by measuring correlation among tuples in the published data. It shows that if correlation values are obtained from published table, it specifies certain amount of privacy disclosure that is present in the current researche on the published dataset that is of "significant" effectiveness. Kifer (2009) demonstrated the effect of attacker based on correlation values from sanitized dataset and the possibilities of stimulating to analyze the results of privacy in a well-organized manner. All these tasks mentioned earlier focus mainly on the privacy protection of published data on the basis of Tuples correlation and doesn't defend privacy against FD, FFD-attack. Data correlation hiding difficulty for publishing data has been evaluated to overcome by researchers (Tao *et al*., 2010). They specifically make use of i masking function to guarantee the level of privacy for each attributes by measuring the correlation relationship among attributes which require hiding perceptive information perfectly. It doesn't defend confidentiality against FFD-attack and CFD-attack for the published micro table

CFD based cleaning methods are more efficient and suitable in defending privacy against CFD-attacks, it

automatically mines CFD patterns from micro data table or original input samples data, it is used to clean CFD based rules in the published data. It is not always easy to design CFDs patterns and it requires long labor-intensive procedures to perform this cleaning process. However, the CFD based cleaning rule detection technique is not suitable to data quality tools. The expansion of FDs is well-known as CFDs and recently, it has been used to distinguish data variation problems which presents a framework that is dependent on cleaning result through SQL (Cong *et al*., 2007). Some of the existing researchess (Zhang *et al*., 2007) investigated the problem of privacy preservation based on the discovery of information. By using these anonymization techniques, intruders easily identify the information regarding authorized user information. This weakness has been presented in the research and this weakness has been overcome by use of the algorithm-based discovery (Jin *et al*., 2010).

Some of the other studies in the literature also detail and valuate the problem of anonymization without taking in to consideration the precise QI and SA attribute to protect privacy against several numbers of rules (Wang *et al*., 2005) through suppression. In order to overcome these problems, creation of various points of views for optimizing effectiveness were studied in (Yao *et al*., 2005) and the anonymization of various set of views were also studied in (Zhang *et al*., 2007). All these researchs defend privacy against generalization to protect privacy guarantee and don't protect privacy against multiple users through different quasi identifiers against FFD, CFD attacks. In order to overcome these problems, it needs to satisfy privacy by ensuring that more than one privacy rule with multiple sensitive attributes against CFD attacks has been taken into consideration, this problem is solved in this research by using the Normalization function.

The Reflective Process Boyce-Codd Normal Form (BCNF) for multiple subtables and fuzzy binomial distribution for initial partition selection: The ultimate objective of this study is to use the results Review In this research, the difficulty of CFD (Fan *et al*., 2008) based privacy-preserving publishing Microdata model has been analyzed with mined CFD patterns results using the mining algorithm (Li *et al*., 2007b). To effectively mine important CFD patterns, in the research done earlier, an improved FP-growth procedure is developed which extracts frequent patterns for Conditional Functional Dependency (CFD) with two scans that are needed for each data instance D over the relation to attain an extremely compacted frequency-descending tree structure (FP-tree). In the research done earlier, a CFD based (d, l) Inference Model is presented to overcome the problem of CFD and Full Functional Dependency (FFD) attack but the major problem occurs when mined CFD patterns published data result in violation of privacy results, if the published data is separated into several number of many sub tables. To conquer this privacy destruction difficulty, a new Normalization method has been proposed for mined CFD patterns that satisfy (d, l) Inference Models.

From mined CFD patterns based published Microdata table, a Boyce-Codd Normal Form (BCNF) is applied to identify privacy violation problems and then it decomposes the violated data table into two to eliminate privacy violation problems. In this study, a new fuzzy binomial frequency distribution is presented to group the partition data from the result (d, l) Inference Model along with the changed publishing data from BCNF. Proposed anonymization algorithm for multiple unsafe CFDs for subtables, measure the privacy accuracy of the system based on both time performance and information loss. The proposed research not only analyses the problem of privacy measures but also addresses the problems of privacy violation between single and multiple subtables.

MATERIALS AND METHODS

Consider a database be microdata table that stores the confidential information about a set of persons. The database D consists of two major attributes: Quasi-Identifiers (QI) and Sensitive Attributes (SA) whose combination can contribute as being the individual key and their confidential information has been represented in Table 1 of the hospital patient discharge. Let's consider the patient discharge data relation can be relating to customers through attributes Country (CO), Area Code (AC), Hospital(H), Age(A), Gender(G), Zip code(ZC), Race(R) and disease (D). Where Race (R) and ICD-9-CM correspond to sensitive attributes(SA) and other attributes correspond to Quasi Identifiers (QI). To overcome the limitation of FFD, motivation of CFD with (d, l) Inference Model in our research, there are some examples that have been studied.

Before going to the proposed framework first CFD has to be defined along with the rule of CFD patterns, from that analysis the best patterns for (d, l) Inference Models against CFD attacks can be determined. A CFD $\varphi = X \rightarrow Y, t_p$ over relation R is assumed to be not

Table 1: Transaction table with CFD of hospital patient discharge data

Transaction	CFD	Patterns
CO,H,R,D	CO,H→R,Disease	{o1|02, 111111, .|.}
CO,AC,H,R	CO,AC,H→R	{01, 023,-, White}
CO,AC,G,D	CO,AC,G→Disease	{01,., Alcohalism|flu}
CO,AC,ZC,D	CO,ZC→Disease	{0245, ., disease}
CO,AC,A,D	CO,AC, A→Disease	{01, 023, 39|54, Disease|flu}
CO,A,D	CO,A→Disease	{01|02, 37-45|48-56, white}

important if Y∈X. If φ is insignificant, either it is satisfied by each one of data 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 data instances in the tuple t such that $t[X] \leq t_p[X]$ (e.g., if $t_p[Y_L]$ and $t_p[Y_R]$)). Each and every CFD patterns that corresponds to the relation r is defined by supp X→Y, t_p, r. It becomes hard to analyze the results of CFD patterns for a larger dataset.

In order to overcomes these problems and semantically measure the results of CFD patterns, first define the semantically related CFD patterns in the following manner and mine important CFD patterns from that patterns using Compact Frequent Pattern Growth Branch Sort algorithm (CFPGBS) method:

$$\phi_0 : \left([CO, ZC] \rightarrow Disease(01, \_ \| \_)\right) \quad (1)$$

$$\phi_1 : \left([CO, H] \rightarrow Disease(01,023 \| Diabetes)\right) \quad (2)$$

$$\phi_2 : \left([CO, H] \rightarrow R(02,40 \| black)\right) \quad (3)$$

$$\phi_3 : \left([CO, H] \rightarrow Disease(01,022 \| HIV)\right) \quad (4)$$

In ,ϕ_0, disease (01, ‖_)) denotes the CFD pattern tuple that enforces a building of semantically related constants for attributes CO, ZC, disease for each tuple t. It states that for customers in India, ZC uniquely determines disease. The privacy lessens if it is in FFD that only holds the subset of Tuples only "CO = 01", rather than onsideration of relation r_0. CFD ϕ_2 assures that for any person in India (country code 01) with Area Code (AC), 023, the disease of the customer must be diabetes as forced by its pattern tuple disease (01, 023‖Diabetes). correspondingly ϕ_2. These cannot be expressed in FFDs. After the CFD are applied to Table 1 the changed variables are given as a new table as in Table 2.

Best CFD patterns are mined from above Table 1 in previous research with pattern mining. Table 2 shows the results of mined CFD based on (d, l) Inference Model,

Table 2: Transaction table after CFD with removal of gender from 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

however it still violates the privacy preservation rules since the AC (Area code) of American contains only one value which makes identifying Diabetes disease easy to identify so the individual privacy of data lessens when compared to all the data privacy. It is required to improve privacy results of (d, l) Inference Model against CFD and FD attacks. In order to overcome these problems first, we categorize the CFD published data table into many subtables and provide privacy rules for each subtable that consists of quasi identifiers and sensitive attributes separately for each subtable.

The main motivation for this study is the observation of the multiple privacy rule against mined CFD patterns in (d, l) Inference Model. The mined CFD patterns based published microdata table is split into multiple subtables with possibly overlapping attributes in such a manner that number of privacy rules can be applied to each sub-table and it is anonymized using (d, l) Inference Model in Table 2. In Table 2, the data is split into multiple subtables based on BCNF function if and only non-trivial CFD patterns do not exist for each attribute on anything other than the superset of the candidate key. Defined is the privacy rules R for each subtable and these do not depend on attribute values for each table against (d, l) Inference Model.

The privacy rules R need a relation r in the Boyce-Codd Normal Form (BCNF) (if and only every determinant is a candidate key and it does not depend on CFD patterns, changes have been shown as below):

$$CC, H \rightarrow R, ICD-9-CM \quad (5)$$

$$H, A \rightarrow ICD-9-CM \quad (6)$$

$$H, ZC \rightarrow ICD-9-CM \quad (7)$$

where, CC, H, A, ZC-candidate keys and R, ICD-9-CM-super keys. Table 2 is decomposed into multiple subtables and it satisfies privacy against (d, l) Inference Model, based on the privacy rules as mentioned above. In this research, we use the following general steps to confirm that the published microdata subtables BCNFT*satisfy every privacy rule.

The privacy rule of BCNF in which A_m may not appear is taken in to consideration. The values of each and every attribute are essentially independent of each other and must satisfy the l-diversity Inference Model in order to protect privacy against CFD patterns. Privacy rule r can be conditioned from the set of privacy rules R if and only it satisfies privacy rule r.

Table 3: Country table for hospital data

Quasi Identifier (QI)		
CO	Area code	Zipcode
India (01)	*	7****
India (01)	*	7****
India (01)	*	7****
India (01)	*	7****
America (02)	*	7****
America (02)	*	7 ****
India(01)	*	7****

Table 4: Hospital table with rule 1

Zipcode	H	Race
7****	111111	Asian
7****	111111	White
7****	222222	Black
7****	111111	White
7****	222222	Black
7 ****	111111	White
7****	333333	White

Table 3 both country code and area code under the category of zip code, violate the privacy rules. In order to resolve this problem instead of specifying country code and area code separately, we access both of these attributes based on the zip code attributes itself by applying privacy rule 1, CC, H→R to Table 3. Now, Table 2 changes as Table 4 without considering the direct relationship between quasi identifiers and sensitive attributes using privacy rules in BCNF. It changes in the following manner and CO, AC codes rows automatically are accessed based on the zip code function. Table 4 satisfies, the BCNF without non trivial CFD patterns for rule 1-3, it has to decompose the original CFD mined patterns into BCNF form:

$$H, A \rightarrow ICD-9-CM \quad (8)$$

$$H, ZC \rightarrow ICD-9-CM \quad (9)$$

Table 3 is again split into multiple subtables along with privacy conditions rule 2 and 3. Table 5 and 6 separately satisfy the privacy rule of two and three for (d, l) Inference Model .Table 4-6 are required to satisfy privacy rules as mentioned above and are automatically grouped into a single table based on the separation of attributes across published tables. Thus, it becomes important to measure whether the specified privacy rule is applicable to all published data tables and multiple subtables too. Next defined as such is the criterion as in BCNF schema for the published microdata table's results in Table 2. Here, it needs to satisfy two conditions for each privacy rule and measure the results of privacy for published microdata table.

Table 5: Hospital table with rule 2

A	H	ICD-9-CM
≤50	111111	HIV
≤50	111111	Diabetes
≤50	222222	Flu
≤50	111111	Diabetes
>50	222222	Diabetes
>50	111111	Diabetes
>50	333333	Flu

Table 6: Hospital table with rule 3

ZC	H	ICD-9-CM
7****	111111	HIV
7****	111111	Diabetes
7****	222222	Flu
7****	111111	Diabetes
7****	222222	Diabetes
7 ****	111111	Diabetes
7****	333333	Flu

- A singular case of <Q1, SA> non-reachability from each other
- A generic case of table in the form of BCNF. The results from BCNF are graphically represented as G = (V, E)

Where each Vertex V in the Graph G corresponds to an attribute such as Quasi Identifiers (QI) and Sensitive Attributes (SA) in Table 4-6. An Edge E in the graph exists among the relationship between two vertices in the graph and it is required to satisfy at least one published table BCNF that contains both attributes such as Sensitive Attributes (SA) and Quasi Identifiers (QI). For example, assuming that Table 7a and b are the results of published data table against (d, l) Inference Model; their results are graphically represented as Table 8c. It shows that the two attributes in the graph such as age and race are not reachable from each other directly, since the attributes are self-regulating from known published tables and satisfies (d, l) Inference Model property.

Consider an example Table 7a and b be the nonsingular case of privacy preserved data based on (d, l) Inference Model against Conditional Functional Dependency (CFD) where Quasi Identifier (QI) and Sensitive Attributes (SA) are easily accessible from each other and violate privacy preservation rule. Since, it satisfies privacy preservation based on (d, l) Inference Model with a single rule privacy rule at a time. In order to overcome these problems, it has to satisfy Q1→SA with multiple sensitive attributes based on (d, l) Inference Model against CFD patterns. Consider an example; Table 7a and b are considered a s a singular case to perform (d, l) Inference Model against Conditional Functional Dependency (CFD) with sensitive attributes and quasi identifiers.

In order to perform this process Table 7b serves as the BCNF for rule 2 H, A→ICD-9-CM where $Ql \cap A_i^* \cup Ql^* =$

Table 7: Published data with BCNF

ZC	R	H	A	ICD-9-CM	Figure
7****	Asian	111111	≤50	HIVs	
7****	White	111111	≤50	Diabetes	
7****	Black	222222	≤50	Flu	
7****	White	111111	≤50	Diabetes	
7****	Black	222222	>50	Diabetes	
7 ****	White	111111	>50	Diabetes	
7****	White	333333	>50	Flu	

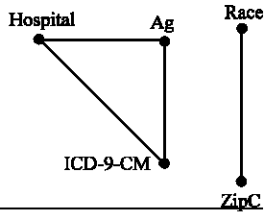

Table 8: American adult dataset attributes information

Attribute name	Attribute information
Age (A)	Continuous
Work Class (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

{H, A}∪{S}and S={ICD-9-CM} A_i*is R and ZC attributes Ql* = {l diversity results attributes}. It shows proposed BCNF based results against CFD in Ql¬SA framework with privacy rules R, moreover Ql and SA cannot be easily reachable from each other in BCNF and it doesn't violate privacy rule R.

Algorithm A; BCNF for CFD attacks

Input: Mined CFD table (CFDT) and privacy rules (R)
Output: Published BCNFT*

Choose an arbitrary privacy rule and anonymize CFDT to BCNFT* in order to enforce it; BCNFT*←{BCNFT*}
Find privacy rule Ql¬SA which violates BCNF over BCNFT*
Compute non CFDT and Remove SA from CFDT
While (not false) do
if (there is a schema R_i in result that is not in BCNF
from table BCNFT* then begin
let Ql¬SA be a non trivial conditional functional dependency that holds privacy rules R_i such that Ql¬R_i is not in non
CFDT and Ql∩SA = ϕ
result: = (results-R_i) ∪ (R_i-SA) ∪ (Ql-SA);
End go to step 2
Else
True

Fuzzy binominal frequency distribution: In this study, a new fuzzy binominal frequency distribution model to group the partition data of (d, l) Inference Model with BCNF results has been presented. In order to analyze the privacy preservation results of BCNF for each and every attribute, it is necessary to deal with the condition of the d closeness frequency value for each attribute in the published table. Consider that S_1 and S be the data attribute values that are close to d closeness frequency distribution results of uniform and skew distribution methods, if $|C_1-C_2| \leq d$ where C_1 and C_2 denote the total number of occurrences of condition that are related to S_1 and S_2.

Consider a group G that contains set of l sensitive data values which are close to frequency distribution values such as uniform and skew distribution, if for any two values S_i, $S_j \in G$ d-close, Known set of l distinct values $S\{S_1, ..., S_n\}$, a partition schema of set S is denoted as $\{P_1, ..., P_t\}$ by segmenting the published data into several groups. Construction of the initial partitions for the bottom-up approach driven is performed by using Fuzzy Binomial Distribution (FBD) for published Microdata table results from BCNF with n independent results such as success/failure that are close to d closeness value with a data table of size n.

The major problem of this research is that it takes results of initial partition frequency that have been calculated based on success or failure rate only. It is not easy to apply this process to all published microdata table. In order to overcome these problems a fuzzy parameter is introduced to analyze results depth manner for initial partition in the bottom up approach. Consider, a credibility theory which provides the information about fuzzy variables and how it is applicable to binominal distribution function to bottom up partition. Let Θ be a

nonempty set and P be the set of data attributes values such as S_1 and S_2 that are close to d closeness frequency distribution results of uniform and skew distribution methods, if $|C_1\text{-}C_2| \leq d$ where C_1 and C_2 denotes the total number of occurrences of condition that are related to S_1 and S_2.

For every S_1 and S_2 that are d-close if $|C_1\text{-}C_2| \leq d$ the event is denoted as d related number is denoted by Cr {d} which indicates the credibility that c will be close to other variables in the BCNF function that occur in the same group G. As per the credibility theory the following three axioms are accepted to perform initial partition of bottom up approach:

- Axiom 1: (Normality) Cr (Θ) = 1
- Axiom 2: (Monotonicity) Cr (d_1)≤Cr (d) whenever $d_1 \subset d$
- Axiom 3: (Self duality) Cr (d_1)+Cr (d_1^c) = 1

A fuzzy variable is used to perform initial partition of the bottom up approach based on a set of attributes (Θ, P, Cr) and it is measured by:

$$\mu(G) = \left\{2Cr\{\xi = G\}\right\} \wedge 1, G \in d \tag{10}$$

In this research, the concept of hybrid binominal frequency distribution for initial partition of bottom up approach has been used. Let ξ be a discrete random variable that corresponds to both Sensitive Attributes (SA) and Quasi Identifiers (QI) with probability of success θ that satisfies d closeness measure for each group G in bottom up approach with triangular membership function.

The Probability Density Function (PDF) of Sensitive Attributes (SA) and Quasi Identifiers (QI) with fuzzy membership function of θ are correspondingly specified by:

$$\Phi(X) = \begin{cases} \binom{n}{x}\theta^x(1-0)^{n-x} \text{ if } X = 0.1...n \\ 0 \text{otherwise} \end{cases} \tag{11}$$

X be number of samples in the BCNF table results:

$$\mu(\theta) = \begin{cases} \dfrac{\theta - a}{b - a} \text{if } a \leq \theta \leq b \dfrac{b - \theta}{c - b} \text{if } b \leq \theta \leq c \\ 0 \text{otherwise} \end{cases} \tag{12}$$

Hybrid binomial distribution for initial partition of quasi identifiers and sensitive attributes are computed using the expression:

$$ch(\xi = r) = \begin{cases} \underbrace{\sup}_{a \leq \theta \leq b}\left\{\min_{\theta}\left(\dfrac{\mu(\theta)}{2}\right) \wedge \binom{n}{r}\theta^r(1-\theta)^{n-r}\right\} \\ \text{if } \underbrace{\sup}_{a \leq \theta \leq b}\left\{\min_{\theta}\left(\dfrac{\mu(\theta)}{2}\right) \wedge \binom{n}{r}\theta^r(1-\theta)^{n-r}\right\} \langle 0 \\ 1 = \underbrace{\sup}_{\theta \in \theta}\left\{\min_{\theta}\left(\dfrac{\mu(\theta)}{2}\right) \wedge 1 - \binom{n}{r}\theta^r(1-\theta)^{n-r}\right. \\ \text{if } \underbrace{\sup}_{a \leq \theta \leq b}\left\{\min_{\theta}\left(\dfrac{\mu(\theta)}{2}\right) \wedge \binom{n}{r}\theta^r(1-\theta)^{n-r}\right\} \geq 0 \end{cases} \tag{13}$$

Bottom-up approach: In the studies and research carried out previously what has been developed is an efficient bottom-up approach to protect the privacy for classified information. The researching principle of bottom up approach is just opposite to the top-down approach. First, the set of unique values S {S_1, ..., S_n} results from BCNF are separated into multiple numbers of groups based on d closeness measure with set of l distinct values. For each partition, S {S_1, ..., S_n}, calculation of the number of removed Tuples by applying frequency distribution of Fuzzy Binomial Distribution (FBD) P is done. Then starting from the initial segmentation, merging takes place for two similar adjacent ones probability distribution as $P_i' = P_i \cup P_{i+1}$ and it can be compared to the information loss IL (P_i') with IL (P_i)+IL (P_{i+1}). If IL (P_i')≥IL (P_i)+IL (P_{i+1}) that will not be merged. Otherwise (P_i) and (P_{i+1}) is merged into single group and construct IL (P_i'). Repeat these steps until there no IL P_i can be merged, it can be finally stored as best frequency distribution as $FBDG_{best}$ which always construct (d, l) inference groups. In bottom up approach based FBD partitions should assure (d, l) Inference Model against all intermediate and final partitions.

Algorithm:

Bottom up construction for (d, l) inference groups
While P probability distribution does not satisfy the (d, l) Inference Model
For all (d, l) Inference Model group G from BCNF table do
Compute information loss among two frequency distribution functions IL (P_i') from fuzzy binominal distribution function
End for
Find the best frequency distribution as $FBDG_{best}$
Perform probability P by $FBDG_{best}$
End while;
Output best frequency occurrence result of bottom up approach

RESULTS AND DISCUSSION

In this study the performance of proposed multiple subtables are measured on the basis privacy preserving methods for Conditional Functional Dependency (CFD)

Table 9: Adult dataset samples

Numbers	Detailed
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, 7-8th, 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

against (d, l) Inference Model with Boyce-Codd Normal Form (BCNF) and compare results with existing mined CFD results. Experimentation results were evaluated for both d, l parameters where d belongs to uniform and skew data frequency distribution against CFD attacks for generalized data. The efficiency of proposed bottom up approach based on fuzzy binomial distribution function and existing bottom up approach based on Log-Skew distribution is compared with each other. Experimentation of proposed and existing methods is carried out using CENSUS dataset which consists of personal information of individuals. CENSUS dataset contains the personal information of the real dataset. Each of the Census dataset's size is 100K.

Census dataset totally consists of 15 attributes and attributes information are mentioned in Table 9. Recorded samples of each American dataset samples have also been mentioned in Table 9. In this research, 1500 samples have been considered as input for privacy preserving (d, l) Inference Model. The attributes mentioned in Table 6 majorly comprise of five attributes such as age, education, marital status, education-num and research class and these are 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 consists of six attributes which are closer to distribution methods such as uniform distribution and skewed distribution. These distribution methods are separately represented as the U-dis dataset and S-dis dataset. It shows the privacy results of proposed bottom up approach with Fuzzy Binomial Distribution (FBD) which are comparatively much faster than the existing

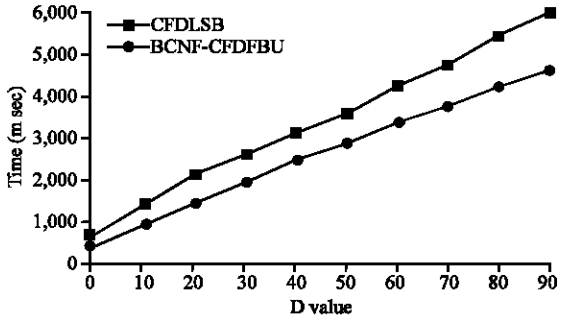

Fig. 1: Time performance comparison U distribution

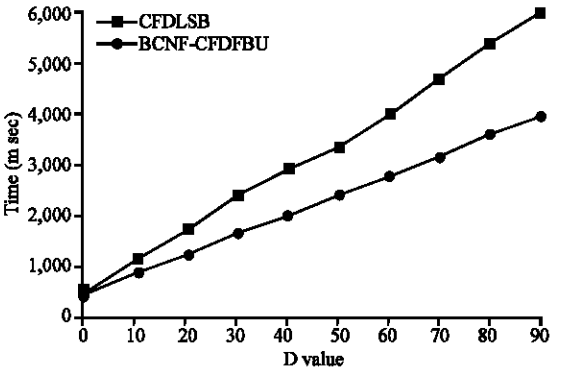

Fig. 2: Time performance comparison S distribution

bottom up approach with log-skew-Normal based frequency distribution, because of the multiple subtable privacy for CFD results. The Time performance results of U and S distribution (BCNF-CFDFBU: Boyce-Codd Normal Form Conditional Functional Dependency Fuzzy Bottom up, CFDLSBU: Conditional Functional Dependency Log-Skew-Bottom Up) as shown in Fig. 1 and 2, respectively.

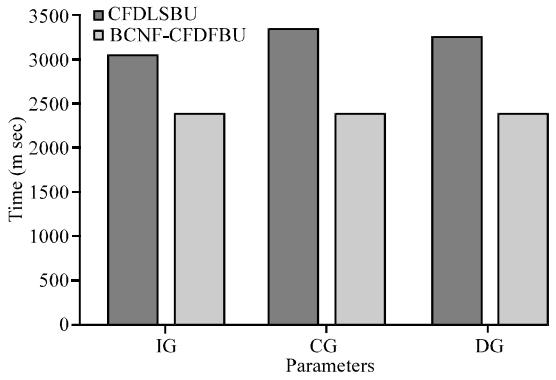

Fig. 3: Time performance comparison, CENSUS dataset for CFD, Parameters of grouping

Figure 3 presents the results of time comparison of grouping methods LG, CG and DG (BCNF-CFDFBU: Boyce-Codd Normal Form Conditional Functional Dependency Fuzzy Bottom up, CFDLSBU: Conditional Functional Dependency Log-Skew- Bottom up) It shows that the time performance of CG is always faster for BCNF-CFDFBU than other CFDLSBU.

The discernbility metric is one of the important privacy preservation metrics used to determine the result of privacy in research that has been previously carried out (Bayardo and Agrawal, 2005). In this metric, every anonymized tuple result from BCNF is allocated to a penalty of term with size of group $|G|$ while a suppressed tuple results from BCNF is allocated to a penalty of the size of the census dataset. The information loss of entire dataset that corresponds to existing and proposed partition methods are predicted using Eq. 14:

$$IL = \sum_{i=1}^{t} \left|G_i\right|^2 + b^* \left|D\right| \quad (14)$$

where, t the total number of groups is results from bottom up approach against (d, l) Inference Model and b is the number of tuple that are removed from each group in bottom up approach. Comparison has been carried of the information loss of our bottom-up approaches with the FBD for BCNF-CFD to existing CFDLSBU results in Fig. 4. It shows that proposed bottom-up approaches with FBD for BCNF-CFDs always return less information loss than existing CFDLSBU.

The information loss comparison result of varied distinct attributes values along with CFD for existing and proposed bottom up approaches results is shown in Fig. 5a, b. Figure 5a, b shows the results of CFD for U-dis and S-dis dataset, It must satisfy l values for CFD that are

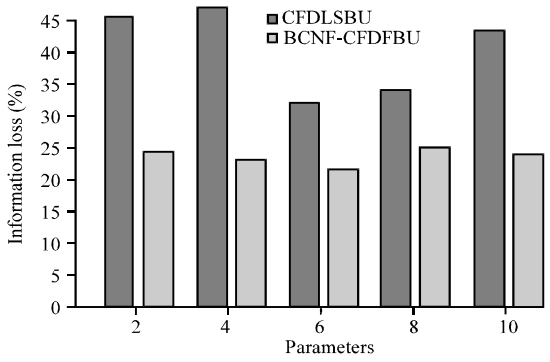

Fig. 4: Information loss comparison for census dataset

closer to d-closeness if the value of l becomes larger the loss of information also high when compared to existing CFDLSBU partition methods proposed BCNF-CFDFBU partition methods have less information loss.

Additionally, use the relative error (Xiao and Tao, 2006) to measure the result of accuracy based on number of queries given by user in the following form:

SELECT (*) FROM Dataset
WHERE pred (A_i), ..., pred (A_{qd})

where, qd is the query measurement value for each query and pred (A_i) denotes a set accurate value of domain values A_i. The size of the entire pred (A_i) set is calculated by the percentage P. Let Act and Est represent the actual query results and estimated query results that corresponds to microdata table T and the published Tables T*, correspondingly. The relative error is defined as follows:

$$\frac{\left|Act - Est\right|}{Act} \quad (15)$$

Figure 6 shows the relative error results among number of privacy rules between conditional functional dependency for BCNF-CFDFBU (Boyce-Codd Normal Form Conditional Functional Dependency Fuzzy Bottom up) and CFDLSBU (Conditional Functional Dependency Log-Skew-Bottom up), it shows that proposed BCNF-CFDFBU have less information loss while using BCNF for QI-SA in CFD.

Figure 7 shows the results of relative error of Uniform distribution among number of privacy rules with BCNF-CFDFBU (Boyce-Codd Normal Form Conditional Functional Dependency-Fuzzy Bottom up) and CFDLSBU (Conditional Functional Dependency Log-Skew- Bottom up), it shows that proposed BCNF-CFDFBU has a

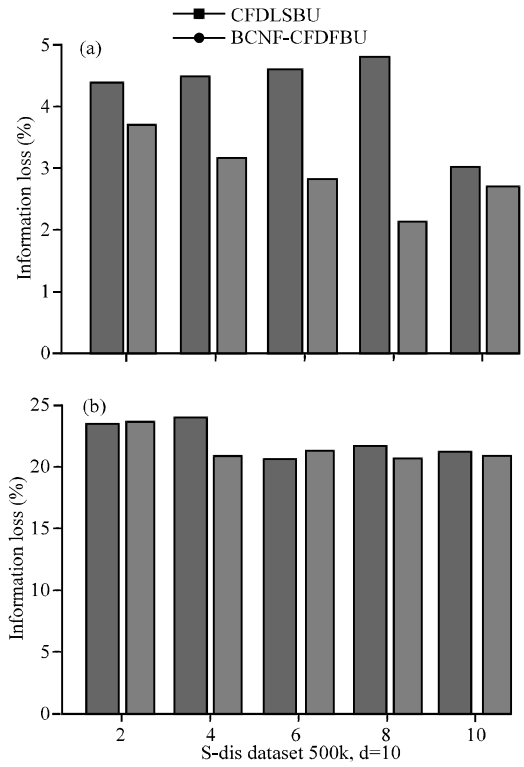

Fig. 5: a) Information loss comparison for U-dis dataset; b) Information loss comparison for S-dis dataset

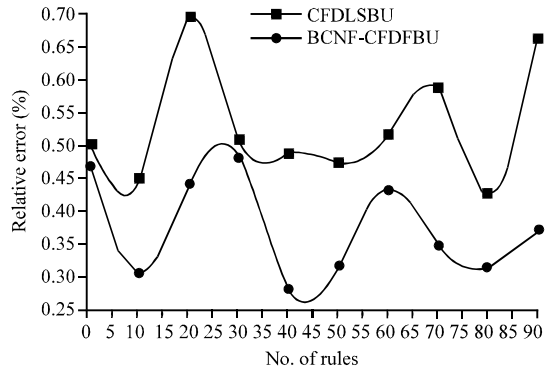

Fig. 6: Relative error on census dataset

less relative error rate when BCNF is used for QI-SA in CFD. Figure 8 shows the results of relative error of skew distribution among number of privacy rules with BCNF-CFDFBU (Boyce-Codd Normal Form Conditional Functional Dependency-Fuzzy Bottom up) and CFDLSBU (Conditional Functional Dependency Log-Skew-Bottom up), it shows that proposed BCNF-CFDFBU has less relative error rate when BCNF is used for QI-SA in CFD.

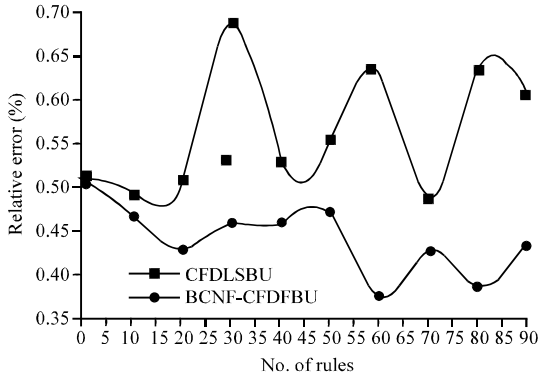

Fig. 7: Relative error on U-dis census dataset

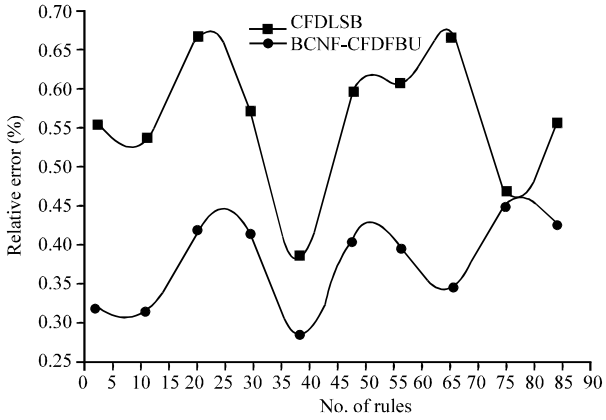

Fig. 8: Relative error on S-dis census dataset

CONCLUSION

The problems of privacy-preserving publishing data for Conditional Functional Dependencies (CFD) against (d, I) Inference Model and the specification of multiple subtables based privacy protection is discussed in this research. To perform multiple subtable privacy preserving publishing for mined CFD against (d, I) Inference Model, a Boyce-Codd Normal Form (BCNF) has been proposed to satisfy privacy rules for all published sub-tables. It accurately defines the (d, I) Inference Privacy Model, in order to protect sensitive information that is caused by CFDs and FFD. The construction of initial partition by using frequency distribution with Fuzzy Binomial Distribution (FBD) based bottom up approach performs well in an efficient manner than the Log-Skew-Normal Alpha-Power Distribution (LSKNAPD) bottom up approach. Our empirical studies are evaluated using census datasets, it is demonstrated that privacy accuracy of Boyce-Codd Normal Form (BCNF) for multiple sub

tables privacy preservation of (d, I) Inference Model against CFD have higher privacy accuracy than existing methods.

REFERENCES

Ali, H., A. Fatlawi, A. Basheer, A. Assadi and M.H. Jabardi, 2015. Dynamic database schema for hospital management system. Asian J. Inf. Technol., 14: 122-128.

Bayardo, R.J. and R. Agrawal, 2005. Data privacy through optimal k-anonymization. Proceedings of the 21st International Conference on Data Engineering ICDE 2005, April 5-8, 2005, IEEE, London, USA., pp: 217-228.

Cong, G., W. Fan, F. Geerts, X. Jia and S. Ma, 2007. Improving data quality: Consistency and accuracy. 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: 315-326.

Fan, W., F. Geerts, X. Jia and A. Kementsietsidis, 2008. Conditional functional dependencies for capturing data inconsistencies. ACM Trans. Database Syst., Vol. 33, No. 2. 10.1145/1366102.1366103.

Jin, X., N. Zhang and G. Das, 2010. Algorithm-safe privacy-preserving data publishing. Proceedings of the 13th International Conference on Extending Database Technology, March 22-26, 2010, ACM, New York, USA., ISBN: 978-1-60558-945-9, pp: 633-644.

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.

Li, J., G. Liu and L. Wong, 2007a. Mining statistically important equivalence classes and delta-discriminative emerging patterns. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 12-15, 2007, ACM, New York, USA., ISBN: 978-1-59593-609-7, pp: 430-439.

Li, N., T. Li and S. Venkatasubramanian, 2007b. T-closeness: Privacy beyond k-anonymity and l-diversity. 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: 106-115.

Machanavajjhala, A. and J. Gehrke, 2006. On the efficiency of checking perfect privacy. Proceedings of the Twenty-Fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems, June 27-29, 2006, ACM, New York, USA., ISBN: 1-59593-318-2, pp: 163-172.

Machanavajjhala, A., D. Kifer, J. Gehrke and M. Venkitasubramaniam, 2007. l-diversity: Privacy beyond k-anonymity. ACM. Trans. Knowl. Discovery Data, Vol. 1 10.1145/1217299.1217302.

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.

Miklau, G. and D. Suciu, 2004. A formal analysis of information disclosure in data exchange. Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, June 13-18, 2004, ACM, New York, USA., pp: 575-586.

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.

Sweeney, L., 2002. k-Anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Based Syst., 10: 557-570.

Tao, Y., J. Pei, J. Li, X. Xiao and K. Yi *et al.*, 2010. Correlation hiding by independence masking. Proceedings of the 2010 IEEE 26th International Conference on Data Engineering (ICDE), March 1-6, 2010, IEEE, Long Beach, CA., ISBN: 978-1-4244-5445-7, pp: 964-967.

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

Wang, K., B. Fung and P.S. Yu, 2005. Template-based privacy preservation in classification problems. Proceedings of the 5th IEEE International Conference on Data Mining, November 27-30, 2005, IEEE Computer Society, Washington, DC., USA., pp: 466-473.

Wang, K., P.S. Yu and S. Chakraborty, 2004. Bottom-up generalization: A data mining solution to privacy protection. Proceedings of the 4th IEEE International Conference on Data Mining, November 1-4, 2004, IEEE Computer Society, Brighton, UK., pp: 249-256.

Xiao, X. and Y. Tao, 2006. Anatomy: Simple and effective privacy preservation. Proceedings of the 32nd VLDB International Conference on Very Large Data Bases, September 12-15, 2006, Seoul, Korea, pp: 139-150.

Yao, C., X.S. Wang and S. Jajodia, 2005. Checking for k-anonymity violation by views. Proceedings of the 31st International Conference on Very Large Data Bases, October 4-6, 2005, VLDB Endowment, Trento, Italy, ISBN:1-59593-154-6, pp: 910-921.

Zhang, L., S. Jajodia and A. Brodsky, 2007. Information disclosure under realistic assumptions: Privacy versus optimality. Proceedings of the 14th ACM Conference on Computer and Communications Security, October 29-November 02, 2007, ACM, New York, USA., ISBN: 978-1-59593-703-2, pp: 573-583.