Volume 5, No. 3, March 2014

Journal of Global Research in Computer Science

ISSN-2229-371X

REVIEW ARTICLE

Available Online at www.jgrcs.info

Detection and Prevention of Leaks in Anonymized Datasets

*Sandeep Varma Nadimpalli and Valli Kumari Vatsavayi

Department of Computer Science and Systems Engineering, Andhra University
Visakhapatnam, Andhra Pradesh, India -530 003
*snvarma9@gmail.com, vallikumari@gmail.com

Abstract: With a wide spread of modern technology, person specificdata dissemination has beenincreasing rapidly,leading to a global concern for preserving privacy of an individual. Several principles like *k*-anonymity, *l*-diversity etc., have been proposed to protect the person specific information during data publishing. However, the presence of dependencies in an anonymized dataset may identify the individual due to the hypothetical nature of the adversary/attacker. This paper shows how the presence of these dependencies among Quasi-Identifiers (QI), Sensitive (S) attributes and also between QI and S attributes can lead to the potential identification of an individual using Bayesian Networks. A solution Break-Merge (BM) was proposed on the fly to reduce the attacker's inferring nature on the sensitive data. Experimentations show the efficacy of theproposed approaches.

Keywords: Preserving Data Publishing, Data Dependencies, Knowledge Breach Attack, Verification, Knowledge Breach Probability, Bayesian Networks.

INTRODUCTION

Rapid growth in hardware technology in memory and storage management increased the storage of high volumes of data. Organizations both public and private are making their data available electronically to enable data access services the World Wide Web. This may lead to disclosing of information to the private/external parties who may use the data for survey or mining purpose. These organizations data may contain person specific data which may be extremely sensitive. Also, this sensitive data may contain possible dependencies that can open the inference channels for the attacker to extract sensitive information easily. Hence the urge and insight of the researchers has begun towards the privacy issues and their struggle for inventing new principles and frameworks for strong privacy protection came into existence.

In USA, when public voter's registration list is combined with the health insurance information records the medical record of the governor of Massachusetts has been potentially identified [1]. This problem was termed as linking attack in the literature [6]. Consider the microdata holding the information of the census information shown in Table I. The *sensitive attributes* in the dataset are *Government, Marital-status and Salary*. These attributes are considered to be private by the individual [1].

The attribute *Name* is termed as *explicit identifier* because one can easily identify the exact tuple by knowing the name of the individual.

For instance, it is clear from Table I that Alice works in a private organization. *Age, Gender, Zipcode attributes* are called as *quasi-identifiers(QI)* because when they are combined with an external dataset there might be a possible potential leakage of the identity of an individual. To prevent this generalization technique (replacing more specific value to less specific value or to replace the value with * termed to be suppression) has been widely used to partition the QI group in the microdata.

k-anonymization is a popular generalization technique used to protect privacy of individuals from the data records. In this technique, the identifying attributes such as *SSN, Name* in the original data table are removed and then the individuals are formed into groups having size of *k* or more [1][3][5]. The 5-anonymized version of the census data is shown in Table II. However, when *k* is large the adversary can infer the sensitive value information with high level of probability[5]. It is shown in [6] that *k-anonymity* does not provide privacy. They define the anonymized table in such a way that for each quasi-identifier group, at most 1/*l* sensitive values must be present.

Motivation

Consider the sample anonymized adult dataset shown in Table II which satisfies 5-anonymity. Form Table II it is clear that the probability to identify an individual is at most 1/5. If an adversary, say Alice knows that Jessica is a female and belongs to the first anonymity-group then Alice can infer that the salary of Jessica is '≤ 50K' with probability of 4/5. If Alice has additional background knowledge about Jessica that she works in a 'State-gov' he can easily identify the marital status and the salary of Jessica with a probability of 1, i.e., 100% inference. The inference **"Sate-gov→Never-married, ≤ 50K"** exists even after the table is anonymized. This inference is termed as quasi- identifier to sensitive attributes association. It is also clear that there is a dependency among quasi, sensitive and quasi to sensitive attributes. In general it is hard to find out how and exactly what attributes are dependent on each other and to what extent just by looking into a large dataset after anonymization.

This being a prime motto we showed how these dependencies between quasi-identifier, sensitive attributes and on both can be discovered by constructing a belief network [41][42] . We term this dependency attack as *Knowledge Breach attack (*KBA). This attack is modeled using Bayesian Network to identify potential inferences in an anonymized dataset.

Table I. Sample Census Dataset.

Explicit Identifier	*Quasi-Identifier*			*Sensitive attribute*		
Name	*Age*	*Gender*	*Zipcode*	*Government*	*Marital-Status*	*Salary*
Alice	90	M	27000	Private	Married-civ-spouse	>50k
Flynn	30	F	18000	State-gov	Never-Married	≤50k
Adam	83	M	26000	Self-emp-inc	Married-civ-spouse	≤50k
Jessica	32	F	13000	Federal-gov	Married-civ-spouse	≤50k
Bob	51	M	58000	Private	Married-civ-spouse	>50k
Calvin	65	M	24000	Private	Divorced	≤50k
June	41	F	23000	Private	Divorced	≤50k
Jane	32	F	16000	Local-gov	Separated	>50k
Scott	73	M	37000	Federal-gov	Never Married	≤50k
Lousy	50	F	22000	State-gov	Never-Married	≤50k

Table II : 5- Anonymized Census Dataset.

Quasi-Identifier			*Sensitive attribute*		
Age	*Gender*	*Zipcode*	*Government*	*Marital-Status*	*Salary*
[30-50]	F	[13000-23000]	State-gov	Never-Married	≤50k
[30-50]	F	[13000-23000]	State-gov	Never-Married	≤50k
[30-50]	F	[13000-23000]	Federal-gov	Married-civ-spouse	≤50k
[30-50]	F	[13000-23000]	Private	Divorced	≤50k
[30-50]	F	[13000-23000]	Local-gov	Separated	>50k
[51-90]	M	[24000-58000]	Private	Married-civ-spouse	>50k
[51-90]	M	[24000-58000]	Self-emp-not-inc	Married-civ-spouse	≤50k
[51-90]	M	[24000-58000]	Private	Married-civ-spouse	>50k
[51-90]	M	[24000-58000]	Private	Divorced	≤50k
[51-90]	M	[24000-58000]	Federal-gov	Never-Married	≤50k

RELATED WORK

Several privacy principle techniques were present in the literature. We divide them categorically and present in this section.

Anonymization Operations

Many of the privacy preserving principles adopt generalization as a basic operation to anonymize the microdata. Three flavors of generalization operations are suppression[7],single-dimension generalization[8][9][10][11][12] and multiple domain generalization [13][14][15]. In suppression technique the QI values in each QI-group are replaced with stars ('*') where as in single-dimension technique, disjoint sub-domains of QI are formed in such a way that each QI value in the microdata is mapped to the other sub domain that contains the corresponding values. For example, Table II shows a single dimension generalization satisfying 5-anonymity. To be more specific, the domain Gender is divided into two sub domains "M" and "F". Multiple domain generalization is an extension of single dimension generalization. Here, the QI values are mapped to the overlapping sub-domains.

Off-the-shelf software's like SAS [24], SPSS [25]and STATA [26]employ suppression and single dimensional generalization techniques. Here, the advantage is the patterns can be easily generated with the help of those tools. Multi-dimensional approach suffers from query analysis, say for instance classification. To this reason, off-the-shelf-software's don't use multi-dimensional technique for statistical analysis instead they adopt suppression so that it can be treated as a missing value and can be processed easily by these statistical software's. Also µ-Args [20] and datafly[22] use suppression and single-domain generalization techniques.

Anonymization Techniques

1)Generalization based Techniques: Statstical community used randomization methods to protecting the privacy of an individual [16]. Also, they added noise to the data before the data release to preserve privacy in fraud detection. However, these methods failed in providing effificient anonymity solution and as a result it lead to data integrity failures. Sweeney proposed *k*-anonymity to protect privacy of the traget individuals in the microdata. The linking of external voters-list with medical data revealed the identity of the personnel [1]. To prevent this linking attack, generalization based techniques are developed [1] [8] [17] [18][19].

k-anonymity techniques fails when the adversary has potential knowledge on the sensitive attributes. To aid this technique *l-diversity*was developed to protect against the inferences on the sensitive values [6]. Later (*α, k*)-anonymity a combined version of *k-anonymity* and *l-diversity* was proposed by Wong et al [11]. It protects both identification and sensitive information by reducing the homogeneity attack. The parameter *α* defines the maximum percentage of any sensitive value within any Q_{id}-block. Another rigid form of *l-diversity* is *m*-variance technique which divides the group such that the group must have exactly m-sensitive values [21]. (*c, k*) safety [23] assumes a stronger background knowledge. If the attacker know *k* pieces of knowledge (*c,k*) safety guarantees the inference of sensitive values adhering to *c* confidence. Techniques like *t*-closeness [2] and (*k,e*)-anonymity [27] can deal with only with numerical attributes while other principles can handle categorical data also. Apart from the above principles Xia and Tao define *personalized privacy* where the individual is give an option to define his/her own degree of privacy level [28].

2) Permutation based Techniques: In these techniques data perturbation is absent. They publish the QI values directly Anatomy [29], *k*-permutation [27], bucketization [23] and ambiguity [30] techniques fall under this category. These methods help in achieving better utility than generalization based methods. Permutation based methods restrict the association of a single sensitive attribute with the respondent's quasi-identifier. However, this can disclose the information when the adversarial knowledge increases potentially [13].

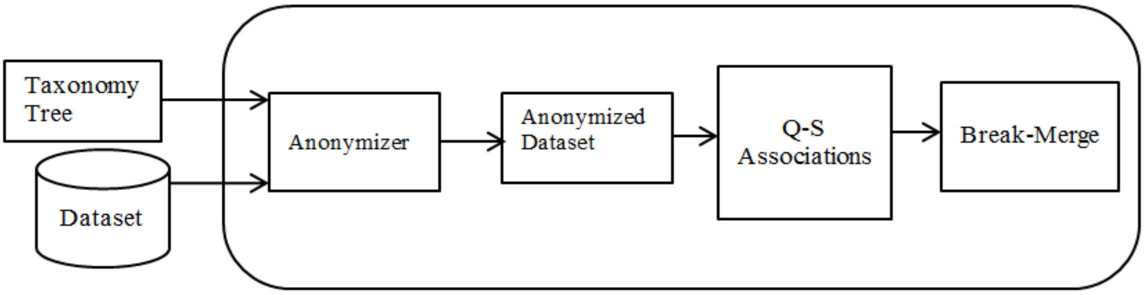

Figure 1.Architecture.

3) Query based Privacy: In query based privacy different view of the dataset is projected. There could be a chance to reveal the private information when different views are merged [31] [32] [33] [34]. [32] [33] provide methods for better privacy and utility. They treat the sensitive values as independent in the released data. However, these techniques areNP-hard [34] [35]. Dwork proposed the concept of differential privacy [36] [37] and [38] extended the work of Dwork.

To summarize, the anonymization technique like *k-anonymity* fails in discovering the associations (background knowledge) among the QI and sensitive attributes. Even *l-diversity* cannot handle these associations when the sensitive attributes in the microdata increase. To prevent these associations Weijia and Shangteng proposed a Q-S association hiding algorithm by using association and disassociation rules and then generalize the sensitive values [39]. For determining the association rules, they construct 1-itemsets using inverted file data structure. However, in practical scenarios the sensitive attributes must not be generalized from utility perspective while extracting useful patterns from the dataset.

In this paper, Bayesian network based model was proposed for detecting dependencies [45] in an anonymized dataset. Also, a solutionis proposed for publishing the dataset such that an adversary cannot re-identify the individual [46].

ARCHITECTURE

The architecture of the proposed method is shown in Figure1. Initially a taxonomy tree and the dataset are fed to the anonymizer for anonymizing the data. After the dataset is anonymized, Bayesian networkis constructed for quasi-identifiers, sensitive attributes individually and a Bayesian net on the whole dataset to detect the dependencies in the anonymized data. Whenfound,Break-Merge technique is used to reduce the plausible inferences in the anonymized dataset. The proposed methods for detecting the dependencies and preventing the dependencies are explained in the subsequent sections.

DEPENDENCIES DETECTION

Researchers collect non-aggregate data from various organizations. However, the adversary with his/her hypothetical nature may have access to various external datasets like voters/medical list for mapping the individuals so that he/she can identify the individuals potentially. Various types of attacks and their remedies were discussed in related work. However, none of those methods focused on how the adversary could potentially identify an individual in an anonymized data. The motivating example clearly shows a dependency existence in an anonymized dataset.

Typically, dependencies among attributes express how well the attribute values are dependent on each other. One example is the SSN number where in at most one individual is associated with one unique SSN number. To detect these kinds of dependencies Bayesian network is employed [41]. In essence, a Bayes net is a directed acyclic graph (DAG) whose edges represent statistical dependencies. However, there may be conditional dependencies among nodes. Each attributes in the dataset is considered as node in the Bayes net (Definition 1). The ICS [44] search algorithm is adopted to preserve the dependency between variables into causal relationship among the nodes/attributes.

Definition 1: *(Bayesian Network)*:

Let $t_i = \{t_{i_1}, t_{i_2}, \ldots\ldots.t_{i_n}\}$ be a tuple, where $t_{i_j}'s$ are the instances of the tuple t_i for given set of N attributes ie., $t_i[X_j] = t_{i_j}\forall$ i & j. The Bayesian Net (BN) is as an ordered pair BN= (G, H) where G is a DAG constructed by using the nodes in the network.

Here the attributes are nodes. More formally if N_i 1≤ i≤ |N| are the set of attributes which are considered as 'n' nodes of the DAG and their edges represent the dependencies between those attributes. H represents the hypothesis on the DAG which is given formally as below.

$$h_{t_i.[X_j]}/_{\pi_i} = P_{BN}(t_i.[X_j]/\pi_i) \qquad (1)$$

where π_i are the set of parents of X_i in G. Hence for the Bayesian network BN discrete probability distribution of the form is defined as follows From (2) it is clear that if $\pi_i = \{\emptyset\}$ i.e., X_i has no parents and hence the distribution is unconditional. Otherwise it is a conditional distribution where in this distribution is expressed in conditional probability tables CPT (see Table 3).

Table III. CPT for ZipcodeAge→Gender dependency.

ZipcodeAge		Gender	
		P_F	P_M
13000-23000	**30-50**	**1**	0
13000-23000	**51-90**	**0.5**	**0.5**
22000-58000	**30-50**	**0.5**	**0.5**
22000-58000	**51-90**	0	**1**

$$P_B(X_1, X_2, \dots\dots X_n) = \prod_{i=1}^{n} P_B({}^{X_i}/_{\pi_i}) = \prod_{i=1}^{n} h_{t_i.[X_j]}/_{\pi_i} \qquad (2)$$

Definition 2 *(Conditional Probability)*:

The Bayesian network BN defines a discrete probability distribution defined in equation (2). The conditional Probability table (CPT) is given by

$P_{BN}(X_i = t_i.[X_i]/X_k = t_k.[X_k]\ for\ X_k \in \pi_i)$Where$\pi_i$are the set of parents ofX_i

$$= \prod_{j=1}^{n} \frac{\text{Freq}(t_j.[X_j])}{\sum_{X_k \in \pi_i} \text{Freq}(t_k.[X_k])} \qquad (3)$$

In the context of privacy the following three cases may arise.

i. The dependencies among quasi-identifiers (QI).
ii. The dependencies among sensitive attributes (S)
iii.The dependencies between QI and S attributes.

The following sections describe each of the case mentioned above with an illustrative example.

Dependencies among Quasi Identifiers

For simplicity the anonymized Table II is used as an illustrating example throughout this paper to show to what level the dependencies are available among quasi-identifiers. Initially a Bayes network structure is formed as shown in the Figure 2. The dependencies are checked on the network diagram. The network clearly shows that *age* is an independent attribute. The attributes *zipcode is dependent on age* and *Gender* is dependent on both age and zipcode. After identifying independent and dependent nodes conditional probability tables are calculated.

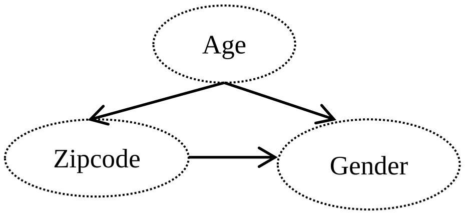

Figure 2.Bayesian Net for Quasi-Identifiers.

For simplicity if the plausible threshold value is between [0.5, 0.75) the risk level is considered to be high. However the privacy risk level can vary from publisher to publisher. The dependencies{Age→Gender,Age→zipcode, Zipcode→Gender and ZipcodeAge→Gender} can hold well for the quasi-identifiers. However all these dependencies might not hold true for the given threshold. So, the conditional probability tables for each of the dependencies are calculated and verified for valid dependencies. Here the dependencies {Age→Zipcode and ZipcodeAge→Gender} holds good.

Table IV. CPT for Age →Zipcode dependency.

Age	*Zipcode*	
	$P_{13000\text{-}23000}$	$P_{22000\text{-}58000}$
30-50	1	0
51-90	0	1

The CPT's for the dependencies in QI attributes are shown in Table III and Table IV. The shaded values show that the dependencies with high probabilities. A privacy threshold is defined to assess how stronger the dependencies in the dataset. The plausible threshold levels were defined based on the conditional probability values as shown in the Table V.This is purely determined by the data publisher and can be changed according to the organization needs.

Table V. Risk Levels.

Score Range	Plausible Privacy Risk Levels (α)
[0.00, 0.20)	Low
[0.20, 0.50)	Moderate
[0.50,0.75)	High
[0.75,1)	Very High

Conditional Probabilities calculation

Example:

For two attributes according to (4) the CPT is given as

CPT (Attribute1=X→Attribute2=Y)

$$\left[= \frac{Count(\ Attribute1 = X) + Count(\ Attribute2 = Y)}{Count\ (Attribute2 = Y)}\right]$$

For instance, the CPT (Age=*(30-50)*→zipcode=*(13000-23000)*)= [5/5] =1. Similarly the remaining values are calculated. The corresponding CPTs for the dependencies Age→Zipcode and ZipcodeAge→Genderare shown in Table III and Table IV.

When we look at the conditional probability Table III for age and zipcode the probability of revealing the age=30-50 is 100% when the adversary knows the zipcode to be 13000-23000. Further, if the adversary knows the zipcode and age values, the probability of finding whether he belongs to M or F group is 50% in some cases and 100% in some cases.

Dependencies among Sensitive attributes

As seen in the first case, dependencies among the quasi-identifiers were identified. In this section,the dependencies among sensitive group are identified. From Figure 3 it can be observed that the Government is independent and the

Table VIII.CPT for Age,Gender,Zipcode,Marital-status,Salary→Government Dependency.

AgeGenderZipcodeMarital-statusSalary					Government				
					$P_{State\text{-}gov}$	$P_{Fedreal\text{-}gov}$	$P_{Private}$	$P_{Local\text{-}gov}$	$P_{Self\text{-}emp\text{-}not\text{-}inc}$
30-50	F	13000-23000	Never-Married	≤50K	**1**	0	0	0	0
30-50	F	13000-23000	Married	≤50K	0	**1**	0	0	0
30-50	F	13000-23000	Separated	>50K	0	0	0	**1**	0
51-90	M	24000-58000	Never-Married	≤50K	0	**1**	0	0	0
51-90	M	24000-58000	Divorced	≤50K	0	0	**1**	0	0

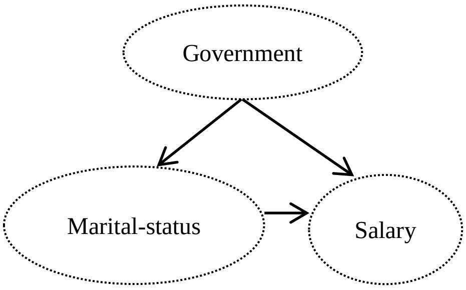

Figure 3. Bayesian Network for the Sensitive Attributes.

remaining attributes Marital-status is dependent on Government and Salary attribute is dependent on both Government and marital status.

The conditional probabilities of the corresponding dependencies are shown in Table VI and Table VII. For instance {GovernmentMarital-status→ Salary} the risk for determining the salary is quite high ($0.5<P_{Salary}<1$) when both Government and marital-status are known to the adversary. The probability for determining Marital-status when Government is known is considerably low. This signifies that the dependency {Government→Marital-status} will not hold in sensitive attributes.

Table VI.CPT for GovernmentMarital-Status⟶SalaryDependency.

GovernmentMarital-status		**Salary**	
		$P_{\le 50k}$	$P_{>50k}$
Stage-gov	Never Married	**1**	0
Federal-gov	Divorced	**0.5**	**0.5**
Private	Divorced	**1**	0
Local-gov	Separated	0	**1**
Private	Married-Civ-Spouse	0	1
Self-Emp-not-inc	Married-Civ-Spouse	1	0
Self-Emp-not-inc	Separated	0.5	0.5

Table VII.CPT for Government⟶Marital-Status Dependency.

Government	**Marital- Status**			
	$P_{Married\text{-}civ\text{-}spouse}$	$P_{Never\ married}$	$P_{Divorced}$	$P_{Widowed}$
State-gov	**0.5**	0.167	0.167	0.167
Federal-gov	0.333	0.333	0.167	0.125
Private	0.125	0.375	0.375	0.125
Local-gov	0.25	0.25	0.25	0.25
Self-emp	0.2	0.4	0.2	0.2

Dependencies among Quasi-Identifiers and Sensitive Attributes

The first two cases show how the dependencies among QI's and Sensitive attribute. Now a Bayesian network for the entire dataset is show in Figure 4. In general the adversary can easily guess the quasi-identifiers since they are grouped and generalized and could be easily identified but the problem arises whenapart from QI, if the adversary has potential background knowledge on sensitive attributes, he checks what dependencies can exist among the attributes by constructing a belief network. From Figure 4, the Government attributes is dependent on the remaining attributes.

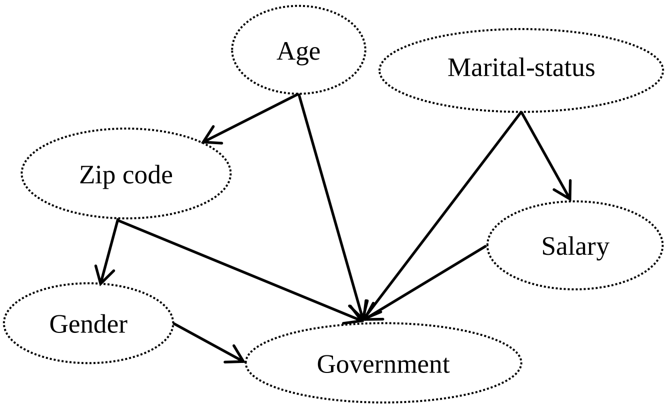

Figure 4. Bayesian Network for the anonymized dataset.

The CPT table for the dependency {Age, Gender, Zipcode, Marita*l*-Status, Salary→Government}as given in Table VIII. It is clear that when an adversary knows the QI group, theMarital-status as never-married and earns a salary <50k he/she can conclude that an individual works in state-gov with likelihood of 100%. So the dependency

{Age, Gender, Zipcode, Marita*l*-Status, Salary→Government} hold good for Table II.

Algorithm

Algorithm 1 detects the dependencies in an anonymized dataset. The original dataset D as shown in the Table I is anonymized to DS* as shown in Table II. Initially, the algorithm assumes that there are no dependencies in the dataset. A Bayesian network is constructed for DS* using ICS search algorithm [44].

From the Bayesian net a set of dependent attributes(N) and independent attributes ($N_i.C$) is found. Now the conditional probability tables CPT for each dependent attribute are calculated. If for each value of N there exist distinct values of independent attribute $N_i.C$ with respect to N with probability less than risk level α (regarding to existing values in the database instances) add the dependency rule $N_i.C$→Ni to DEP (steps 2-7). Now examine each edge linking from each independent attributes in $N_i.C$. If more than two independent attributes exist for the dependent attributes repeat steps (2-7). If the criterion is satisfied for each and every edge, then add all such dependency rules to DEP set (steps 8-19).

Algorithm 1 : *Detecting Dependencies in an Anonymized Dataset*

```
Input : An Anonymized Dataset (DS*)
Output: Dependencies FD
Assumptions :
N= {$N_1, N_2, \dots\dots. N_n$} // Set of Dependent Attributes
T ={$T_1, T_2, \dots\dots. T_n$}  // Set of conditional probability
tables   for each attribute
$N_i.C$={$C_1, C_2, \dots\dots. C_l$} // be the set of Independent
                              attributesof each  attribute$N_i$
Temp= {$\emptyset$ } // Temporary set for intermediate operations
FD={$\emptyset$ }  // For storing dependencies of the anonymized
dataset

1  Begin
2      For eachattribute$N_i$in N
3        If(|$N_i.C$ |==1) then
4          For each tuple $t_j$ in $T_i$
5            For each attribute $a_k$ in T
6              If($t_j$ [$a_k$]>$\alpha$)
7                FD =FD ∪{$N_i.C \rightarrow N_i$}

8        If(|$N_i.C$ |>1)
9          For each$C_l$ in $N_i.C$
10           For each tuple $t_j$ in $T_i$
11             For each attribute $a_k$ in T
12       If($t_j$ [$a_k$]> $\alpha$)
13             Temp = Temp ∪{$N_i.C_l \rightarrow N_i$}
14       count++;
15       If (count==|$N_i.C$ |)
16             FD =FD ∪{$N_i.C \rightarrow N_i$}
17           Else
18           FD = $FD$ ∪ $T$emp;
19     End For
20 End
```

BREAK-MERGE (BM)

The pervious section showed how dependencies exist in an anonymized dataset. The presences of these dependencies may identify an individual potentially.This may in turn violate the privacy. In order to remove these associations and reduce the attackers' guessing nature of the individual Break-Merge technique is proposed where the anonymized table is separated into Quasi identifier table QIT (Does not hold any sensitive information) and sensitive table ST. Initially QIT is formed in such a way that each group is assigned a Group_Id in a new separate column (Table IX).

The ST's (Tables X, XI, XII) hold marital-status, salary and Government along with the count values of their respective QI-groups are given.The Sensitive tables are represented asset of (*Gid, SA, Count)* where *Gid* is the group id of the corresponding QIDT, SA is the sensitive attribute value and *count* is the number of times the sensitive value is present in the corresponding *Gid* groups respectively. For example, whenthe Government ST is considered, it signifies that the value sate-gov is associated with two tuples in the first QI-group, Federal-gov is associated with one tuple in the first group and one tuple in the second QI-group, Private is associated with one tuple in the first group and three tuples in the second group and so on. In this fashion all the sensitive tables are constructed.

TABLE IX. QI Table

Age	Gender	Zipcode	Group_Id
[30-50]	F	[13000-23000]	1
[30-50]	F	[13000-23000]	1
[30-50]	F	[13000-23000]	1
[30-50]	F	[13000-23000]	1
[30-50]	F	[13000-23000]	1
[51-90]	M	[22000-58000]	2
[51-90]	M	[22000-58000]	2
[51-90]	M	[22000-58000]	2
[51-90]	M	[22000-58000]	2
[51-90]	M	[22000-58000]	2

TABLE X. Sensitive Table of Marital-Status

Group_Id	Marital-Status	Count
1	Married-civ-spouse	1
1	Never-married	2
1	Divorced	1
1	Separated	1
2	Married-civ-spouse	3
2	Never-married	1
2	Divorced	1

With the proposed approach an adversary cannot infer the association between the quasi-identifiers and sensitive attribute because the QIT will not represent any of the information related to the sensitive values and the sensitive values information must be obtained from ST which further increases the probability for the adversary to identify the sensitive value of an individual and hence preserving privacy.

For instance, let us consider that the adversary knows the age of Jessica is 31 and zipcode is 13000 (Tuple id 4 as shown in the figure 1). Since the values in the QIT are generalized and no sensitive information (considering that the adversary wants to know the salary of Jessica) is available.

Table XI. Sensitive Table of Salary

Group_Id	Salary	Count
1	≤ 50k	4
1	>50k	1
2	≤ 50k	3
2	>50k	2

The only information the adversary can guess is the group id i.e., 1, since all the females fall in the first group. With the help of the group id when he looks for ST, he figures out that out of 5 records, 4 females were drawing the salary ≤50K and

1 female person is drawing a salary >50K. From this it is clear that the probability that the salary of Jessica is either 4/5 or 1/5 and hence the adversary must randomly guess from the ST but not the exact tuple.

Table XII. Sensitive Table Of Government

Group_Id	Government	Count
1	Sate-gov	2
1	Federal-gov	1
1	Private	1
1	Local-gov	1
2	Private	3
2	Self-emp-not-inc	1
2	Federal-gov	1

Further the adversary has some prior background knowledge about Jessica that she works in a state-gov. If the adversary wants to know the marital-status and salary of Jessica, The probability to find the marital status of Jessica is 1/2 and probability to find the salary is 1/2 i.e, 50%. Let us extend this much further, even if the adversary knows that Jessica works in a state-gov company the probability that Jessica is never-married and her salary is ≤50K is 2/5 * 4/5 = 8/25 i.e, 32% likelihood or the probability that Jessica is divorced and her salary is >50 will be 1/5 * 1/5 =1/25 i.e., 4% likelihood. From this it is very clear that even if the adversary has sufficient background knowledge on Quasi-Identifier and one of the sensitive attribute, the likelihood for the adversary to infer other sensitive values will be reduced considerably. For simplicity Table XIII presents the probabilities in terms of likelihood for the adversary to guess Jessica's record if the sensitive attribute *'Government'* is known (here knowing that Jessica is working in sate-Government is the background knowledge).

Table XIII. Sensitive Table Of Government

Government	*Marital-status*	*Salary*	*Probability*	*Likelihood*
Sate-gov	Never-married	≤50K	$\frac{2}{5}*\frac{1}{5}$	32%
Sate-gov	Never-married	>50K	$\frac{2}{5}*\frac{1}{5}$	8%
Sate-gov	Married-civ-spouse	≤50K	$\frac{1}{5}*\frac{4}{5}$	16%
Sate-gov	Married-civ-spouse	>50K	$\frac{1}{5}*\frac{1}{5}$	4%
Sate-gov	Divorced	≤50K	$\frac{1}{5}*\frac{4}{5}$	16%
Sate-gov	Divorced	>50K	$\frac{1}{5}*\frac{1}{5}$	4%
Sate-gov	Separated	≤50K	$\frac{1}{5}*\frac{4}{5}$	16%
Sate-gov	Separated	>50K	$\frac{1}{5}*\frac{1}{5}$	4%

The probability breach is defined with the following background knowledge. This is termed as knowledge breach probability (KBP)

Case 1: If the adversary knows only the knowledge on QI group of intentional individual (here Jessica) then the knowledge breach probability is defined as follows.

Definition 3: Given QIT and ST's with n attributes, the group id GID$_l$, the sensitive attribute SA, the knowledge breach probability of an individual when the adversary knows the target individual group id as follows

$$KBP[t.[]\,|t.[n+1]=l]=\prod_{i=1}^{|SA|}\frac{|c_m^j|}{|GID_l|}$$

$$=\frac{\prod_i^{|SA|}|c_i^j|}{|GID_l|^{|SA|}}$$

Where t.[] represents the tuple under all the attributes of the QIT and ST, $t.[n+1]$ represents the n+1th attribute instance and |SA| = No of Sensitive attributes and $|c_m^j|$ is the count of the sensitive value

Case 2:If the adversary has knowledge only on the sensitive valuethe knowledge breach probability is given as follows:

Definition 4: Given QIT and ST having n attributes, the group id GID$_l$, the sensitive attribute SA, the knowledge breach probability of an individual when the adversary knows one of the sensitive attribute value s_m^j i.e., t $[SA_m]$ = s_m^j after reconstructing QIT and ST is as follows,

$$KBP[t.[]|t.[SA_m]=s_m^j]=\frac{|c_m^j|}{|GID_l|}$$

Case 3:If the adversary has the knowledge about QI group and one of the sensitive values of intentional individual then the knowledge breach probability as follows:

Definition 5:Given QIT and ST's having n attributes, the group id GID$_l$ and the sensitive attribute SA. We define the knowledge breach probability of an individual when the adversary knows the group id and one of the sensitive attribute value s_m^j i.e., t $[SA_m]$ = s_m^j of the target individual after reconstructing QIT and ST's is as follows,

$$KBP\big[t.[]\,|t.[n+1]=l\,\&\,t.[SA_m]=s_m^j\big]=\prod_{\substack{i=1\\i\neq m}}^{|SA|}\frac{|c_m^j|}{|GID_l|}$$

$$=\frac{\prod_i^{|SA|}|c_i^j|}{|GID_l|^{|SA|-1}}$$

Lemma 1. When natural join is applied on QIT and ST's the resultant table of n attributes is of the form (t[1],t[2]…t[n],GID,$v_1, v_2, \ldots.\, v_{|SA|}, C_1(v_1), C_1(v_2)\ldots, C_1(v_{|SA|})$) where GID is the group id , v_i is the sensitive value of SA_i and $C_i(v_i)$ is the number of tuples in QI_{GID_l} , where 1≤ i ≤ |SA|. If the adversary having background knowledge on group id and on one of the sensitive values along with their count in the corresponding QI group, the probability to infer the right target tuple will be not less than$\frac{1}{C_i^l(v_i).GID_l|^{|SA|-1}}$, which is given formally as follows.

$$\text{KBP}\big[t.[]\,|t.[n+1]=l\,\&\,t.[SA_m]=s_m^j\big] \geq\frac{1}{C_i^l(v_i).GID_l|^{|SA|-1}} \qquad (4)$$

Proof: A tuple t ∈ ΔT (anonymized table) will be present in any QI group. When we join QIT and ST's the resultant table will contain at least $|GID_l|^{|SA|-1}$ tuples. Since the adversary knows the group id '*l*' and one of the sensitive values along with their count $C_i^l(v_i)$ in the corresponding QI group of the target tuple the possible number of tuples the adversary can guess would be $C_i^l(v_i).\,GID_l|^{|SA|-1}$. Thus the probability that the adversary to infer the right tuple in worst case will result into equation (4). □

Property 1. When an adversary knows the background information of a tuple and when he/she reconstructs QIT and ST's the knowledge breach probability of a tuple when the adversary knows only the QI group is equal to the product of knowledge breach probability of a tuple when the adversary knows only sensitive value and the knowledge breach probability of the tuple when the adversary knows both QI group and Sensitive value of the tuple. Formally,

$$\text{KB}P[t.[]\,|t.[n+1]=l] \\ = \text{KBP}\big[t.[]|t.[n+1]=l\,\&\,t.[SA_m]=s_m^j\big] \\ \text{XKBP}[t.[]|t.[SA_m]=s_m^j]$$

Proof: From the definitions of 3, 4 and 5

$$KBP\big[t.[]|t.[n+1]=l\,\&\,t.[SA_m]=s_m^j\big] \\ \text{XKBP}[t.[]|t.[SA_m]=s_m^j]$$

$$=\frac{\Pi_i^{|SA|}|c_i^j|}{|GID_l|^{|SA|-1}}\ \text{X}\ \frac{|c_m^j|}{|GID_l|}$$

$$=\frac{\Pi_i^{|SA|}|c_i^j|}{|GID_l|^{|SA|}}$$

$$=KBP[t.[]\,|t.[n+1]=l] \qquad \square$$

Corollary 1: The KBP of a tuple when the adversary knows both QI group and a sensitive value of the tuple is always greater than the KBP of a tuple when the adversary knows only QI group. Formally

$$KBP\big[t.[]\,|t.[n+1]=l\,\&\,t.[SA_m]=s_m^j\big] \\ \geq\ \text{KBP}[t.[]\,|t.[n+1]=l]$$

Proof :
Since KBP$[t.[]|t.[SA_m]=s_m^j]$ is ≤ 1 By Property (1) we have

$$KBP\big[t.[]\,|t.[n+1]=l\,\&\,t.[SA_m]=s_m^j\big] \\ \geq\ KBP[t.[]\,|t.[n+1]=l] \qquad \square$$

Property 2: The probability the target tuple t can be revealed by an adversary is always greater than $\frac{1}{|GID_l|^{|SA|}}$ when he knows QI group.

Proof: From the *Property 1* and *Lemma 1* we have

$$\frac{KBP[t.[]\,|t.[n+1]=l]}{KBP[t.[]|t.[SA_m]=s_m^j]}=KBP\big[t.[]\,|t.[n+1]=l\,\&\,t.[SA_m]=s_m^j\big] \\ \geq\ \frac{1}{|c_m^j|\,|GID_l|^{|SA|-1}|}$$

$$KBP[t.[]\,|t.[n+1]=l]\geq\frac{1}{|c_m^j|\,|GID_l|^{|SA|-1}|}\ \text{X}\,\frac{|c_m^j|}{|GID_l|}\ \geq\ \frac{1}{|GID_l|^{|SA|}}$$

$$\square$$

BREAK- MERGE ALGORITHM

The Break-Merge algorithm is as follows. Initially the original dataset DS as shown in the Table I is anonymized i.e., ADT*. For anonymizing the dataset *k*-anonymity and *l*-diversity privacy principles were applied. Without any loss of generality and less information loss (*k, l*) anonymized dataset as shown in the Table II was generated. The anonymized dataset is given as the input to the algorithm as shown in the Figure 2. Initially it is assumed that QIT and ST are empty. It is also assumed that the QIT will contain only quasi identifiers tuples and sensitive table contains only sensitive value data. No other data will be present in both the tables. The algorithm has 2 phases. The first phase is to assign group id to the anonymized data set (line 2-7).

```
Algorithm2: Break-Merge
Input : An Anonymized Dataset (DS*)
Output: Quasi-Identifier Table (QIT) & Sensitive Tables (ST's)
Assumptions :
QIT=Φ, ST_i={ST_1,ST_2,….,ST_m}= Φ,
ADT*= Φ, Gcnt=1;
QIDSet_i={QID1,QID2,…,QID_d}
SASet_j={ SA_1,SA_2,….,SA_m}
1  Begin
2    for each Tuple Ti in DS* do
3      If(QIDSet_i= QIDSet_{i+1}) then
4        Insert Record(QIDset_i, Gcnt, SASet_i)  IntoADT*;
5      else
6        Insert Record (QIDset_i, Gcnt, SASet_i) IntoADT*;
7        Gcnt=Gcnt+1;
8    for i=1 to Gcnt
9        for each tuple T_j $ADT_i^*$
10         Insert tuple (QIDSet_j, i) into QIT;
11     for eachSA_k in SASet{SA_1,SA_2,….,SA_m}
12       for each district SA_k value v in $ADT_i^*$
13           $support_{i\,j}$(v)=The number of record in with
                     $ADT_i^*$sensitive value v;
14         Insert record(i,v,$support_{i\,j}$(v)) into ST_j;
15 End
```

Figure 2. Algorithm for Break-Merge

Initially each tuple in the QI group iscompared with the next tuple in the group and if they match a common group id is assigned to both the tuples. This process is repeated until all thetuples in the anonymized data that forms QI partitions will have their corresponding group-ids. Once the anonymized table ADT* is produced the second phase begins. In this phase the anonymized data is divided into quasi-identifier table (QIT) and sensitive tables (ST's).

For each tuple t_j in ADT_i^* the tuples are inserted into QIT which contains only the QI group associated with its corresponding group id having the form (QIDSet$_j$, i) (step 10) and when coming to the sensitive tables, since *l*-diversity principle is applied on the sensitive attributes, the sensitive values are much diversified keeping *l*-to be large so as to decrease the adversaries inferring attack range. For each QI group the corresponding sensitive values count is calculated and then inserted into sensitive table in the form of (QID, sensitive value (v), count) (step 11-14).

Finally when the QIT, ST's are formed and if the adversary wants to know about any particular individual he/she can reconstruct a tuple by merging QI and sensitive tables using simple natural join. With the proposed algorithm as shown in Algorithm 2 by breaking the anonymized table into QI and sensitive tables the probability breach of the adversary will decrease drastically.

COMPLEXITY

Let n be the number of record and G_{cnt} be the total number of QI groups and S_{cnt} be the total no of sensitive attributes and s_i be the number of diversified sensitive values in each QI group then the total time complexity for splitting the anonymized table into QI and sensitive tables is O (n) + O (G_{cnt} (m + $S_{cnt}(s_i)$)) where m is the number of records in each QI group.

EXPERIMENTATION

Experimentations are conducted in two phases. In the first phase, the scalability is measured for constructing the Bayesian networks. In the second phase experiments were conducted for Break-Merge technique for scalability performances on different real world and synthetic datasets. A comparison analysis is also performed with [39]

Experimentation Setup

1) Phase I

Experiments were conducted on Adult dataset available at UCI machine Learning Repository (UCI). The dataset consists of 14 attributes and 48,842 tuples. The final dataset consists of 30,162 tuples after removing the missing values "?". Out of 14 attributes age, Gender and zipcode were treated as quasi-identifiers and Government, marital status and salary were treated as sensitive attributes.

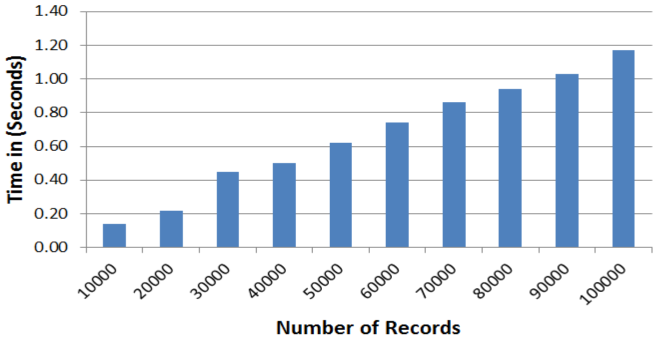

Figure 5. Bayesian net construction time with single sensitive attribute.

Weka tool was used to construct the Bayesian net [45]. Experimentswere conducted on both single sensitive attributes and multiple sensitive attributes. The dataset is replicated such that each equivalence size is 1000. For the construction of Bayes network it took less than 1.2 seconds and nearly 1.7 seconds for single and multiple sensitive attributes respectively (Figure 5 and Figure 6) for a dataset with 1,00,000 records.

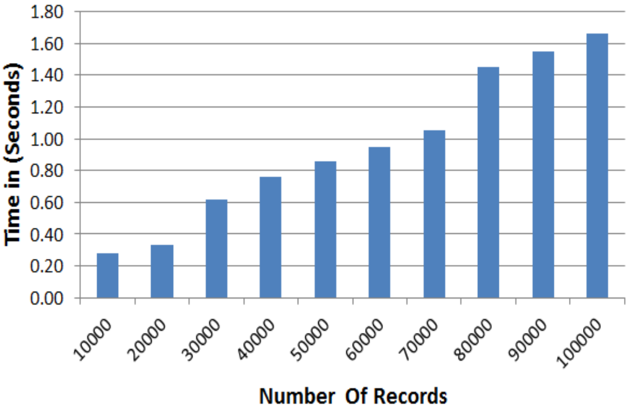

Figure 6. Bayesian net construction time with three sensitiveattributes.

2) Phase II

The experimental setup for break-merge technique is same as in phase I. The Break-Merge algorithm was implemented in Java 1.7 using Netbeans 7.0 IDE. A comparison study was made with [39].

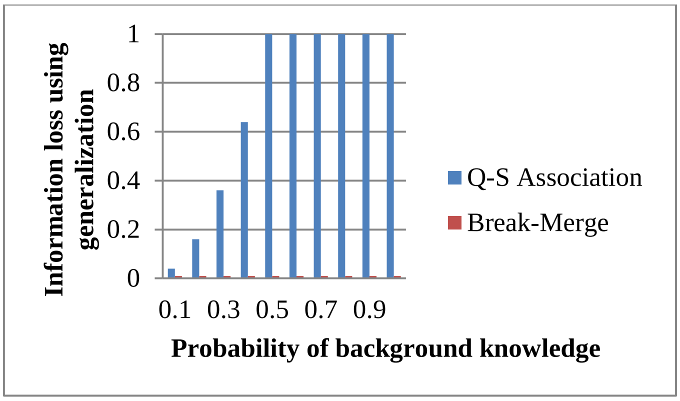

Figure 7.Comparison between Q-S Association and BM.

Weijia et alconstruct 1-itemset to determine the presence of associations between quasi-identifier and sensitive attributes. They construct the rules until they reach the certain threshold. Once the rules were obtained the corresponding sensitive attributes were generalized accordingly.

However, our approach does not generalize the sensitive attributes instead break the table in to QIT and ST's. This increases the utility of the dataset for deriving useful patterns. Figure 7 shows that if the probability of the attacker increases above 50% the sensitive attributes are generalized to a high level and there after remains in that high level in Q-S association but our approach do not generalizing the sensitive attributes the information loss with respect to the sensitive attributes will be zero.

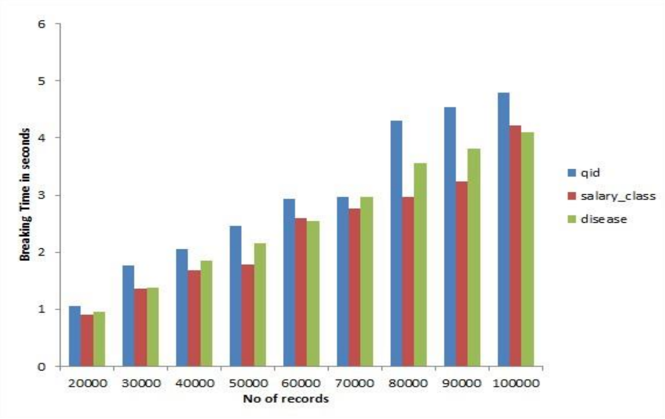

Figure 8. No of records Vs Breaking time with 3 sensitive attributes.

Different performance measures for breaking the tables on real time adult dataset and synthetic dataset that are generated from the adult dataset were done. It took less than 5 seconds to break the dataset that contains 1,00,000 records. The remaining datasets disease and salary also took less than 5 seconds as shown in Figure 9 and Figure 10 respectively.

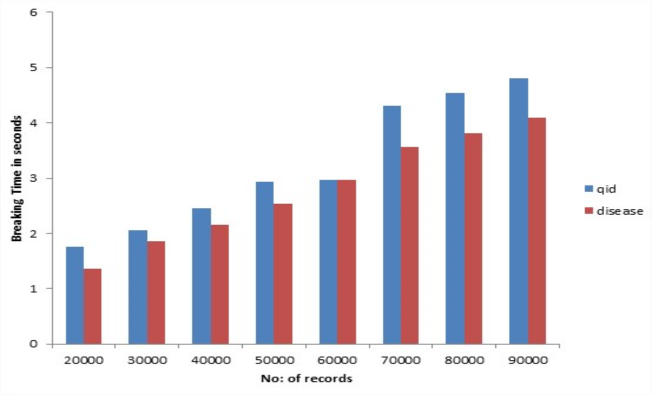

Figure 9. No of records Vs Breaking time for disease dataset.

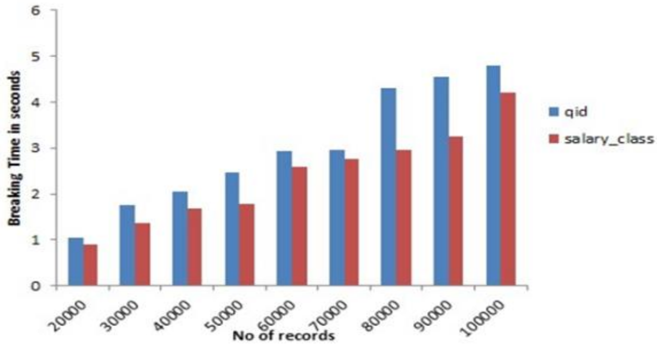

Figure 10. No of records Vs Breaking time for salary dataset.

7. CONCLUSIONS

In this paper, solution for protecting the inferences of sensitive data in an anonymized dataset is discussed. The publisher releases the anonymized data without any verification of vulnerable nature of the adversary on the anonymized dataset. An approach to show how dependencies still prevail in an anonymized dataset using belief networks was discussed.

For this purpose plausible risk levels were defined to show the risk levels in an anonymized dataset. For releasing the data when dependencies exists an approach Break-Merge has been proposed. This approach publishes the data by reducing the attackers inferring nature drastically. This work is of its first kind in the literature addressing the verification techniques for an anonymized datasets. The experimental evaluations show that approaches are practically feasible and scalable.

REFERENCES

[1] L. Sweeny, *k*-Anonymity. A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge based systems, 10(5):557-570, 2002.

[2] N. Li, T. Li and S. Venkatasubramanian, *t*-closeness. Privacy beyond *k*-anonymity and *l*-diversity. In Proc.of ICDE, pages 106-115, 2007.

[3] P. Samarati. Protecting Respondents identities in microdata release, TKDE, 13(6):1010-1027, 2001.

[4] P. Samarati and L. Sweeney. Generalizing Data to Provide Anonymity When Disclosing Information. In Proc. of PODS, 1998.

[5] Meyerson and R. Williams. On the complexity of Optimal*k*-anonymity. In Proc. of PODS, 2004.

[6] Machanavajjhala, J. Gehrke and D. Kifer: *l*-diversity. Privacy beyond *k*- anonymity. In Proc. Of ICDE, 2006.

[7] N. R. Adam and J. C. Wortmann. Security-control methods for statistical databases: A comparative study. ACM Computing Surveys, 21(4):515-556, 1989.

[8] R. Bayardo and R. Agrawal. Data privacy through optimal *k*-anonymization. In Proc. of ICDE, pages 217-228, 2005.

[9] C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In Proc. of ICDE, pages 205-216, 2005.

[10] V. Iyengar: Transforming data to satisfy privacy constraints. In Proc. of SIGKDD, pages 279-288, 2002.

[11] R. C.-W.Wong, J. Li, A. W.-C. Fu, and K. Wang. (α,k)-anonymity: an enhanced *k*-anonymity model for privacy preserving data publishing.In SIGKDD, pages754-759, 2006.

[12] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.W.-C. Fu. Utility-based anonymization using local recoding. In SIGKDD, pages 785-790, 2006.

[13] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis: Fast data anonymization with low information loss. In VLDB, pages 758-769, 2007.

[14] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan: Mondrian multidimensional *k*-anonymity. In ICDE, 2006.

[15] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan: Workload-aware anonymization. In Proc.of KDD, pages 277-286, 2006.

[16] J. Kim:Method for limiting disclosure of microdata based on random noise and transformation. In SurveyResearch Methods of the American Statistical Association, pages 328-387, 2001.

[17] Charu C. Aggarwal: On *k*-Anonymity and the Curse of Dimensionality. In Proc. of VLDB, 2005.

[18] L. Kristen, D. David and R. Ramakrishnan:Mondrian multidimensional *k*-Anonymity. In Proc. of ICDE, 2005.
[19] P. Samarati, L. Sweeney: Generalizing data to provide anonymity when disclosing Information. In Proc. of PODS, 1998.
[20] Hundepool and L. Willenborg: μ-argus: Software for statistical disclosure control. In International Seminaron StatisticalConfidentiality, 1996.
[21] X. Xiao and Y. Tao. *m*-invariance: Towards privacy preserving re-publication of dynamic datasets. InProc. of SIGMOD, pages 689-700, 2007.
[22] L. Sweeney. Datafly: A system for providing anonymity in medical data. In Proc. of DBSec, pages 356-381, 1997.
[23] J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y.Halpern: Worst-case background knowledge for privacy preserving data publishing. In Proc. of ICDE, pages 126-135, 2007.
[24] SAS Institute. SAS/STAT 9.2 User's Guide. SAS Publishing, 1 edition, 2008.
[25] SPSS Inc. SPSS 16.0 Base User's Guide. SPSS Inc., 2 edition, 2007.
[26] Stata Corporation. Stata User's Guide Release 8.0.Stata Press, 1 edition, 2003.
[27] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu:Aggregate query answering on anonymized tables. In Proc. of ICDE, pages 116-125, 2007.
[28] X. Xiao and Y. Tao. Personalized Privacy. In Proc. of SIGMOD, 2006.
[29] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In Proc. of VLDB, pages 139-150, 2006.
[30] Hui W Wendy: Ambiguity: Hide the Presence of Individuals and Their Privacy with Low Information Loss, In Proc. of CSI, India, 2008.
[31] N. Dalvi, G. Miklau, D.Suciu: Asymptotic conditional probabilities for conjunctive query. In Proc. of ICDT, pages 289-305, 2005.
[32] Deustch, Y. Papakonstantinou: Privacy in database publishing. In ICDT, pages 230-245, 2005.
[33] G. Miklau, D. Suciu: A formal analysis of information disclosure in data exchange. In Proc. of SIGMOD,pages 575-586, 2004.
[34] Machanavajjhala, J. Gehrke: On the efficiency of checking perfect privacy. InProcs. of PODS, pages 163-172, 2006.
[35] Barak, K. Chaudhuri, C. Dwork, S.Kale, F.McSherry, K.Talwar,. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In Proc. of SIGACT-SIGMOD-SIGART, Symposium on Principles of Database Systems, pages 273-282, 2007.
[36] Dwork. Differential privacy. In Proc. of ICALP, pages 1-12, 2006.
[37] Dwork, F. McSherry, K. Nissim, A. Smith. Calibrating noise to sensitivity in private data analysis. In Proc. of TCC, pages 265-284, 2006.
[38] K. Nissim, S. Raskhodnikova, A. Smith. Smooth sensitivity and sampling in private data analysis. In Proc. of STOC, pages 75-84, 2007.
[39] Y.Weijia and H. Shangten. *k*-Anonymization without Q-S Associations. In Proc. of WAMI, pages 753-764, 2007.
[40] UCI Repository of Machine Learning databases, http://www.ics.uci.edu/~mlearn/MLRepository.html.
[41] R. E. Neapolitan, Learning Bayesian Networks, Pearson Education, 2004.
[42] Pearl, Probabilistic Reasoning in intelligent Systems: Networks of Plausible Inference, Morgan KaufmannPublishers, San Franciso, CA, USA, 1988.
[43] R. R. Bouckaert, Bayesian Network Classifiers in Wekafor Version 3-5-5, The University of Waikato, 2007.
[44] T. Verma and J. Pearl.An algorithm for deciding if a set of observed independencies has a causalexplanation. In Proc. of Uncertainty in Artificial Intelligence, pages 323-330, 1992
[45] N. SandeepVarma, V. ValliKumari. BM (Break-Merge): An Elegant Approach for Privacy Preserving Data Publishing, In Proc. of PASSAT, IEEE, pages 1202-1207, 2011.
[46] N. SandeepVarma, V. ValliKumari. Detecting Dependencies in an anonymized dataset, In Proc. Of ICACCI, ACM, pages 82-89, 2012.

SHORT BIODATA OF ALL THE AUTHOR

N. SandeepVarma received B.Tech degree in Information Technology from JNTU Hyderabad, Andhra Pradesh, India in 2007. He received his M.Tech in 2009 from Andhra University (Double Degree Program with BTH, Sweden) and is currently pursuing his Ph.D. in Computer Science and Systems Engineering at Andhra University. He also worked as a Senior Research Fellow (SRF) in DST funded project by Department of Science and Technology, Ministry of Science and Technology, Government of India. His research interests include Data Privacy, Formal Verification techniques on security protocols and Software Engineering Software Metrics, Software Testing and Software Estimation models. He is a student member of IEEE.

Dr. V. ValliKumari is currently a Professor in Computer Science and Systems Engineering department and is also Honorary Director of Andhra University Computer Centre. She has over twenty two years of teaching experience. She was awarded a gold medal for the best research in 2008 by Andhra University. Her research areas include Web Mining, Data and Security Engineering and have90 publications in various conferences and journals of international and national repute. She is an active member of IEEE, ACM, CRSI and CSI. She is also the founder vice-chair for IEEE Vizag bay Sub-Section and a fellow of IETE.