Computing and Informatics, Vol. 35, 2016, 1160–1188

VERIFICATION IN PRIVACY PRESERVING DATA PUBLISHING

Sandeep Varma Nadimpalli

Department of Information Science and Engineering
BMS College of Engineering
Bengaluru, Karntaka, India
e-mail: sandeepvarma.ise@bmsce.ac.in

Valli Kumari Vatsavayi

Department of Computer Science and Systems Engineering
AU College of Engineering, Andhra University
Visakhapatnam, Andhra Pradesh, India
e-mail: vallikumari@gmail.com

Abstract. Privacy preserving data publication is a major concern for both the owners of data and the data publishers. Principles like k-anonymity, l-diversity were proposed to reduce privacy violations. On the other side, no studies were found on verification on the anonymized data in terms of adversarial breach and anonymity levels. However, the anonymized data is still prone to attacks due to the presence of dependencies among quasi-identifiers and sensitive attributes. This paper presents a novel framework to detect the existence of those dependencies and a solution to reduce them. The advantages of our approach are i) privacy violations can be detected, ii) the extent of privacy risk can be measured and iii) re-anonymization can be done on vulnerable blocks of data. The work is further extended to show how the adversarial breach knowledge eventually increased when new tuples are added and an on the fly solution to reduce it is discussed. Experimental results are reported and analyzed.

Keywords: Privacy, anonymized data, dependencies, Bayesian net, breach, verification, publishing

1 INTRODUCTION

A wide and enormous growth of information technology in terms of computation and storage, has resulted in collection of large volumes of data by organizations and vendors (Public/Private). This data is made available electronically over the internet for several statistical, data mining, business intelligence and research analysis. Such data may contain individual specific sensitive information for example disease, credit card transaction, etc. When public voter registration list is combined with the health insurance information records [37] the medical record of the governor of Massachusetts has been potentially identified. This problem was termed as linking attack [20]. Principles and frameworks for stronger privacy protection were developed. The k-anonymity is the first known technique to protect the privacy of data by anonymizing it [37, 38]. The main idea of k-anonymity is to make the individuals indistinguishable from others in the published table. This is achieved through generalization such that each individual's tuple has to be same as at least $(k-1)$ other tuples. The concept of generalization technique is to replace more specific value to less specific value, for instance changing the value of age from 23 to [20–25].

Identifier	Quasi-Identifier			Sensitive Attribute		
Name	**Age**	**Gender**	**Zipcode**	**Government**	**Marital-Status**	**Salary**
Alice	90	M	27000	Private	Married	$> 50\,\mathrm{K}$
Flynn	30	F	18000	State-gov	Never-Married	$\leq 50\,\mathrm{K}$
Adam	83	M	26000	Self-emp	Married	$\leq 50\,\mathrm{K}$
Jessica	32	F	13000	Federal-gov	Married	$\leq 50\,\mathrm{K}$
Bob	51	M	58000	Private	Married	$> 50\,\mathrm{K}$
Calvin	65	M	24000	Private	Divorced	$\leq 50\,\mathrm{K}$
June	41	F	23000	Private	Divorced	$\leq 50\,\mathrm{K}$
Jane	32	F	16000	Local-gov	Separated	$> 50\,\mathrm{K}$
Scott	73	M	37000	Federal-gov	Never-Married	$\leq 50\,\mathrm{K}$
Lousy	50	F	22000	State-gov	Never-Married	$\leq 50\,\mathrm{K}$

Table 1. Sample adult data

According to [7] the attributes in the dataset are classified as identifier[1], quasi-identifier (QI), sensitive attributes (S) and Non-Sensitive attributes (NSA). Let us consider the census information in Table 1. Here the sensitive attributes are Government, Marital-status and Salary. These attributes are generally considered to be private by the individual [37]. The attribute *Name* is termed as identifier because one can easily identify the exact tuple by knowing the name of the individual. For instance, it is clear that Flynn works in state-gov. The attributes *Age, Sex, Zipcode* are termed as quasi-identifiers (QI) as these attributes when combined with external dataset like voters registration list may result in potential leakage of the identity of an individual. This identification is due to the presence of common attributes

[1] Identifier sometimes is termed as explicit identifier

among the datasets. Those attributes which do not fall into above three categories are non-sensitive attributes.

Quasi-Identifier			Sensitive Attribute		
Age	**Gender**	**Zipcode**	**Government**	**Marital-Status**	**Salary**
[30–50]	F	[13000–23000]	State-gov	Never-Married	$\leq 50\,\mathrm{K}$
[30–50]	F	[13000–23000]	State-gov	Never-Married	$\leq 50\,\mathrm{K}$
[30–50]	F	[13000–23000]	Federal-gov	Married	$\leq 50\,\mathrm{K}$
[30–50]	F	[13000–23000]	Private	Divorced	$\leq 50\,\mathrm{K}$
[30–50]	F	[13000–23000]	Local-gov	Separated	$> 50\,\mathrm{K}$
[51–90]	M	[24000–58000]	Private	Married	$> 50\,\mathrm{K}$
[51–90]	M	[24000–58000]	Self-emp	Married	$\leq 50\,\mathrm{K}$
[51–90]	M	[24000–58000]	Private	Married	$> 50\,\mathrm{K}$
[51–90]	M	[24000–58000]	Private	Divorced	$\leq 50\,\mathrm{K}$
[51–90]	M	[24000–58000]	Federal-gov	Never-Married	$\leq 50\,\mathrm{K}$

Table 2. 5-anonymized adult data

1.1 Motivation and Our Contributions

Anonymity techniques proposed in the literature [32, 18, 19] observed that due to the presence of dependencies among the attributes in an anonymized dataset an individual can be identified with certain probability. In this section, we show how the background knowledge of the adversary could possibly reveal the individual's identity in an anonymized data.

Consider an anonymized dataset as shown in Table 2. The anonymity level of the dataset is $k = 5$. Here, an individual can be identified with a probability of at most 1/5. If an adversary knows that Bob is male and he belongs to the second anonymity-group then he can infer that the salary of Bob is $> 50\,\mathrm{K}$ with probability of 2/5, i.e. with 40 % likelihood. If an adversary possesses additional background knowledge about Bob, that he works in a private organization one can say that the marital status is Married and he earns salary of $> 50\,\mathrm{K}$ with a probability of 1, i.e. with 100 % likelihood, because the dependency "Private $\rightarrow$ Married, $> 50\,\mathrm{K}$" holds good even though the dataset is anonymized. This inference is termed as *'quasi-identifier to sensitive attributes'* dependency. The dependency attack is said to occur with a probability called as breach probability. For a given scenario, dependency is possible among attributes. Verifying a large dataset manually would be tedious task. Moreover, to find the impact of these dependencies on privacy in the published data is not known. The following research questions need to be answered.

(a) What could be the plausible types of dependencies that can exist among attributes? How to determine the breach probability?

(b) How can we reduce these kind of dependencies?

(c) How can we reduce the breach probability when new tuples are added in the subsequent releases? Is there a limit?

Earlier versions of this work addressed the first two questions [32, 31]. [32] addressed the research question (a) i.e., how the dependencies in an anonymized data can be determined. The conditional probabilities are found using belief network. A belief network or Bayesian network is used to model a domain containing uncertainty [6, 40]. The group of attributes which can identify an individual with highest probability is found. The second research question (b) is addressed in [31]. Break-Merge technique was used to reduce the dependencies. The link between QI and S attributes is de-associated. It is observed that privacy risk increases when new tuples are added in the next release. This paper overcomes this limitation and hence addresses the third research question (c). An incremental Break-Merge technique is proposed in this paper and the increase in privacy breach when new tuples are added is discussed. In addition, the number of the tuples that must be added when a breach threshold is known is also discussed in this paper.

1.2 Organization of the Paper

The rest of the paper is organized as follows: Section 2 discusses the state of art in the literature. Section 3 presents the architecture of the proposed technique. Section 4 discusses the proposed algorithm to find the dependencies among the attributes. Section 5 presents an approach to eliminate the dependencies. A technique to reduce privacy breach for an incremental data release is addressed in section 6. Section 7 presents our experimentation results, and Section 8 concludes the paper.

2 RELATED WORK

Existing privacy techniques use generalization for anonymizing the data. Several other frameworks like suppression [1], single-dimension generalization [4, 12, 15, 43, 47] and multiple domain generalization [13, 18, 19] are the different forms of generalization. In generalization the more specific value is replaced with less specific value whereas in suppression the values are replaced with special characters like '?', '*', etc. In single-dimension technique, every QI value is mapped to the corresponding QI value in the other domain and in multiple domain generalization the QI values are mapped to the overlapped domains. Off-the-shelf softwares like SAS [33], SPSS [34], STATA [35] implement suppression and single dimensional generalization for their statistical analysis as they can be easily processed. These methods do not provide enough information for classification. μ-Args [14] and datafly [36] also use suppression and single-domain generalization techniques.

Statistical community limit their study to randomization techniques for resolving re-identification problem [17]. They added noise to the data but their methods could not provide a better effective solution and therefore lead to failure of data integrity. Observing this data integrity failure researchers moved to generalization

based techniques. [37] proposed the notion of k-anonymity to protect privacy of an individual. However, the adversary knowledge on QI values easily re-identified the individual when the data was combined with outside data. To overcome the linking attack, generalization techniques were developed [38, 4, 2, 29]. Several other techniques like l-diversity [21] were published to protect the inference on the sensitive values. (α, k)-anonymity [43] is the combination of k-anonymity and l-diversity principles. It protects both identification and sensitive information by reducing the homogeneity attack. [46] discusses m-variance as a rigid version of l-diversity which divides the group such that it has exactly m-sensitive values. (c, k)-safety [23] assumes a stronger adversary background knowledge. If the attacker knows k pieces of knowledge (c, k) safety guarantees the inference of sensitive values adhering to c-confidence. t-closeness [20] and (k, e) anonymity [48] deal with only numerical data. l-diversity and (α, k) anonymity handles categorical data. In addition to above, Xia and Tao define personalized privacy [44] where the individual defines his/her own degree of privacy. The primary motivation for this paper is that none of the above works guarantee the claimed anonymization level for the given anonymized data.

Other important methods like anatomy [45], k-permutation [48], bucketization [23], ambiguity [42] help in achieving better utility than generalization based methods but publish the QI values directly. Publishing QI values directly may re-identify the individual when the adversarial knowledge increases potentially [13]. Generalization, permutation and query-based anonymization techniques fail in discovering the associations among the QI and S attributes. [41] proposed Q-S association technique where they generate association and disassociation rules and then generalize the sensitive values. They use inverted file data structure to construct 1-itemsets for determining the association rules effectively. This results in utility loss. All the aforementioned frameworks and principles do not verify the privacy leaks once the dataset is anonymized.

3 ARCHITECTURE OF THE PPDP VERIFICATION MODEL

The architecture of the Privacy Preserving Data Publishing (PPDP) verification model consists of the following four steps, as shown in Figure 1.

Step 1. Anonymizer: Anonymizer uses k-anonymity principle to anonymize the data. This paper generates concept hierarchy tree based on attribute values for a given domain. For example, the concept hierarchy tree for gender could be generalized to person for male or female.

Step 2. Detecting Dependencies: In this step, the dependencies among the QI and sensitive attributes are identified. The attribute dependencies are determined using a Bayesian graph.

Step 3. Break-Merge: To reduce the number of dependencies the dataset is split into QI Tables and Sensitive Tables (ST) separately. This step limits the attackers from inferring knowledge about sensitive data.

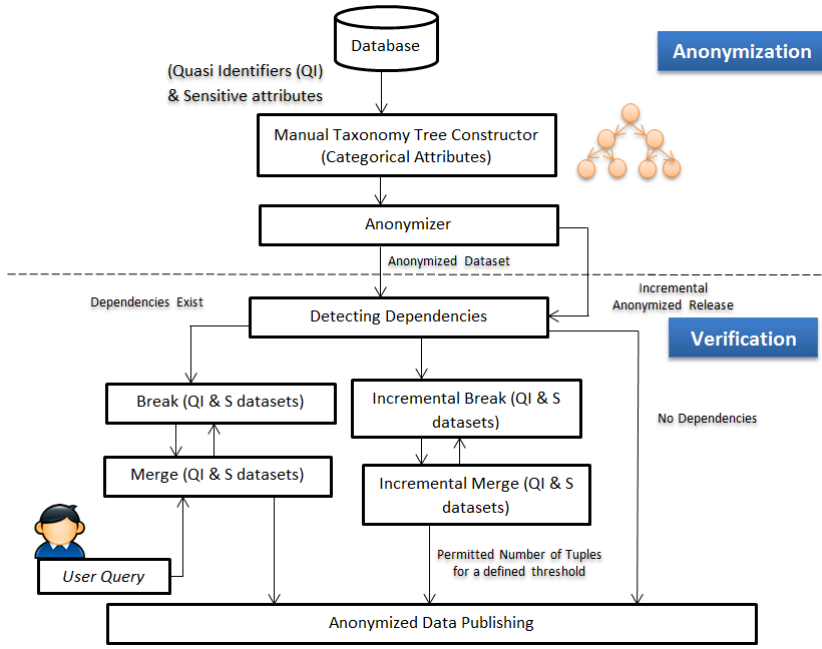

Figure 1. Verification model for privacy preserving data publishing

Step 4. Incremental Break-Merge: This is useful in preserving privacy in subsequent releases of the data. Since the number of tuples vary from release to release Break-Merge is repeated. Addition of new tuples is not permitted if breach threshold condition is violated. This condition helps in determining the limit on the number of tuples if addition of them can make the data vulnerable to privacy attack. If there are no dependencies the anonymized dataset can be published as it is.

The key focus of this paper is to detect attribute dependencies in the context of privacy. Once the dependencies are identified Break-Merge technique is applied to reduce them. However, Break-Merge is useful only for a static release of the data. When new tuples are added for the next release, Incremental Break-Merge is applied to detect dependencies and a suggestion on the threshold, number of tuples that can be added while preserving of the privacy is made. The following sections discuss in detail the proposed verification techniques.

4 DETECTING DEPENDENCIES

The problem of attribute dependencies is modelled using Bayesian network. The Bayesian network (BN) [26] represents a high-level probability distribution on a set

of variables to represent a model on the problem domain given a hypothesis. The parent-child relationship between nodes in a Bayesian network indicates the direction of causality between the corresponding variables. In this paper, the attribute is treated as the node in the Bayesian network.

Definition 1 (Breach Probability)). Let A= $\{A_1, A_2, \ldots, A_n\}$ be the set of QI and S attributes of an anonymized data set D', $t_i = \{t_i.A_1, t_i.A_2, \ldots, t_i.A_n\}$ be a tuple, where $t_i.A_j$ represents the value of the tuple t_i in the possible domain of $A_j \forall j = 1$ to n. Let Π_j be the set of parents of A_j in G. Then the breach probability (BP) for revealing a tuple $t_i.A_j$ known its parents t_j is equal to the Bayesian probability (P_{BN}) of A_j given Π_j. Formally we define it as follows:

$$BP(t_i.A_j/\pi_j) = P_{BN}(t_i.A_j/\pi_j). \tag{1}$$

However, in some cases the entire tuple t_i will be revealed. Now the discrete conditional probability distribution is defined as the product of probabilities of A_j given over Π_j. Formally it is given as

$$BP[t_i.(A_1, \ldots, A_n)] = \prod_{j=1}^{n} BP(A_j/\pi_j) = \prod_{j=1}^{n} H(t_i.A_j/\pi_j) \tag{2}$$

where $H(t_i.A_j)$ is a hypothesis define on $t_i.A_j$. For example the hypothesis could be knowing the values of Age as [20–30]. From (2) it is clear that if $\pi_i = \{\emptyset\}$, i.e., A_j has no parents, the distribution is unconditional else it is conditional distribution. The distribution is calculated in terms of conditional probabilities. Once the Bayesian network is constructed CPT for each node is calculated.

In purview of privacy the following three cases of attribute dependencies may arise.

- The dependencies among quasi-identifiers,
- The dependencies among sensitive attributes and
- The dependencies between quasi identifiers and sensitive attributes.

The following sections discuss each of the cases. The anonymized table (Table 2) is used as an example throughout this paper.

4.1 Dependencies Among Quasi Identifiers

To show the attribute dependency levels among quasi-identifiers a Bayes network is constructed as shown in the Figure 2. The following attribute dependencies are drawn.

- Age is an independent attribute and has no parents.
- Age → Zipcode (Zipcode is dependent on Age)

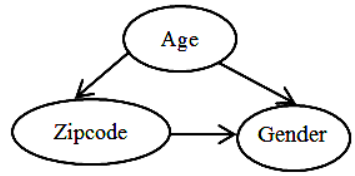

Figure 2. Bayesian net for quasi-identifiers

- AgeZipcode → Gender (Gender is dependent on both Age and Zipcode)

  For two attributes the CPT is given as

$$CPT(Attribute1 = X \rightarrow Attribute2 = Y)$$

$$= \frac{Count(Attribute1 = x) + Count(Attribute2 = y)}{Count(Attribute2 = y)}$$

For the attribute pair Age, Zipcode the CPT(Age = (30–50) → Zipcode = (13000–23000)) = [5/5] = 1. Similarly the remaining values are calculated. The corresponding CPT for the dependency ZipcodeAge → Gender has been shown in Table 3.

It can be seen in CPT(Age, Zipcode) in Table 3 that the probability of revealing that the age is in the range 30–50 is 100 % when the adversary knows the zipcode is in the range of 13000–23000. If the adversary knows the zipcode and also age values the probability of finding whether gender is "F" is 1. In this way the dependencies Age → Gender, Age → Zipcode, Zipcode → Gender and ZipcodeAge → Gender are found. However, which dependencies hold good are determined based on the risk level (α). The (α) value will be varied based on the privacy requirements of the data publisher. For example, if the publisher considers the α value is in between $[0.5, 0.75)$ the dependencies {Age → Zipcode and AgeZipcode → Gender} hold well for quasi-identifiers.

Zipcode	Age	Gender	
		P_F	P_M
[13000–23000]	[30–50]	**1**	0
[13000–23000]	[51–90]	**0.5**	**0.5**
[24000–58000]	[30–50]	**0.5**	**0.5**
[24000–58000]	[51–90]	0	**1**

Table 3. CPT for AgeZipcode → Gender dependency

The dependencies among sensitive attributes follows the quasi-identifiers detection process.

4.2 Dependencies Among Quasi-Identifiers and Sensitive Attributes

Bayesian network on the whole anonymized dataset of Table 2 is shown in Figure 3. Due to the additional background information the adversary may predict by examining the dependencies using a belief network. For example, the government attribute is dependent on the remaining attributes. When the CPT table is checked for the dependency, as given in Figure 4, it can be clearly observed that when the adversary knows the QI value and Marital-Status is Never-Married and earns a Salary $\leq$ 50 K. One can conclude that the individual works in State-gov with likelihood of 100 %. Other combinations also potentially leak the government adversary with complete likelihood. Hence the dependency {Age, Sex, Zipcode, Marital-Status, Salary $\rightarrow$ Government} holds well in the anonymized dataset.

The Algorithm for detecting the dependencies is explained in [32]. To reduce these dependencies [31] proposed a technique Break-Merge which is discussed in Section 5.

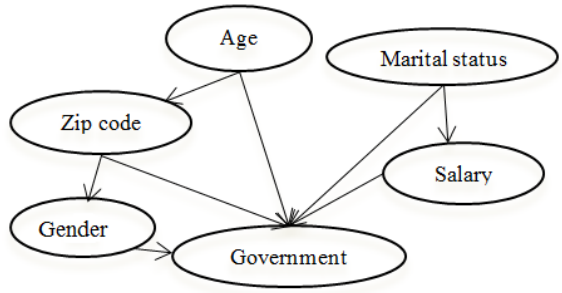

Figure 3. Bayesian net for the anonymized table

Age	Sex	Zipcode	Government						
			$P_{Marital\text{-}Status}$	P_{Salary}	$P_{State\text{-}gov}$	$P_{Federal\text{-}gov}$	$P_{Private}$	$P_{Local\text{-}gov}$	$P_{Self\text{-}emp}$
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 4. CPT for Age, Sex, Zipcode, Marital-Status, Salary $\rightarrow$ Government dependency

5 BREAK-MERGE

The earlier section discussed the presence of attribute dependencies and how one can identify an individual thus violating individual privacy. Earlier versions of this paper proposed Break-Merge [31] technique to reduce the attackers inferring nature. Break-Merge splits the anonymized dataset into quasi-identifier (QI) table and sensitive tables (ST's). The QI table is represented in the form of (QI-Attributes value, G_{ID}) and the Sensitive tables in the form of (G_{ID}, SA, Count) where G_{ID} is the group id of the corresponding QI group, SA is the sensitive attribute value and count is the number of times the sensitive value present in the corresponding G_{ID} groups, respectively. For example, when we consider the Government ST, 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. In this way all the sensitive tables are constructed. Here the QI table does not hold any sensitive information and all the values in the QI are generalized using k-anonymity. Tables 4, 5, 6 and 7 show QI and ST tables.

Age	**Sex**	**Zipcode**	G_{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	[24000–58000]	2
[51–90]	M	[24000–58000]	2
[51–90]	M	[24000–58000]	2
[51–90]	M	[24000–58000]	2
[51–90]	M	[24000–58000]	2

Table 4. QI table

G_{ID}	**Marital-Status**	**Count**
1	Married	1
1	Never-Married	2
1	Divorced	1
1	Separated	1
2	Married	3
2	Never-Married	1
2	Divorced	1

Table 5. $ST_{Marital-Status}$

Since this technique clearly detaches the links between quasi-identifier and sensitive attributes the adversary cannot clearly infer any sort of dependencies between

G_{ID}	**Salary**	**Count**
1	$\leq 50\,\mathrm{K}$	4
1	$> 50\,\mathrm{K}$	1
2	$\leq 50\,\mathrm{K}$	3
2	$> 50\,\mathrm{K}$	2

Table 6. ST_{Salary}

G_{ID}	**Government**	**Count**
1	State-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

Table 7. $ST_{Government}$

QI and S attributes because the QI table does not hold any sensitive information and if needed must be obtained from the ST tables. This will reduce the adversary inferring nature even though he/she possesses information of an individual because the probability for inferring the sensitive value will increase gradually. For instance, let us consider that the adversary knows the age of Bob is 51 and zipcode is 58000 (Tuple id 5 as shown in the Table 1). Since the values in the QI tables are generalized if the adversary wants to know the salary of Bob the only way to guess is through group id i.e., 2, since all the males fall in the second group. With the help of the group id when he looks for ST it is clear that out of 5 records, 3 males are drawing the salary $\leq 50\,\mathrm{K}$ and 2 male are drawing a salary of $> 50\,\mathrm{K}$ and hence the probability that the salary of Bob is either 3/5 or 2/5.

Government	**Marital-Status**	**Salary**	**Probability**	**Likelihood**
State-Gov	Never-Married	$\leq 50\,\mathrm{K}$	$\frac{2}{5} * \frac{4}{5}$	32 %
State-Gov	Never-Married	$> 50\,\mathrm{K}$	$\frac{2}{5} * \frac{1}{5}$	8 %
State-Gov	Married	$\leq 50\,\mathrm{K}$	$\frac{1}{5} * \frac{4}{5}$	16 %
State-Gov	Married	$> 50\,\mathrm{K}$	$\frac{1}{5} * \frac{1}{5}$	4 %
State-Gov	Divorced	$\leq 50\,\mathrm{K}$	$\frac{1}{5} * \frac{4}{5}$	16 %
State-Gov	Divorced	$> 50\,\mathrm{K}$	$\frac{1}{5} * \frac{1}{5}$	4 %
State-Gov	Separated	$\leq 50\,\mathrm{K}$	$\frac{1}{5} * \frac{4}{5}$	16 %
State-Gov	Separated	$> 50\,\mathrm{K}$	$\frac{1}{5} * \frac{1}{5}$	4 %

Table 8. Knowledge breach probability

However, if the adversary gathers much hypothetical information about Bob, let us say, that he works in a private firm and the adversary wants to know the marital-

status and salary of Bob, the probability that Bob is married and his salary is ≥50 K is $2/5 * 3/5 = 6/25$, i.e. 24 % likelihood. This example shows that if the adversary has sufficient knowledge on quasi-identifier and one of the sensitive attribute, the likelihood to infer the remaining sensitive values is reduced considerably. Table 8 presents the likelihood for one combination.

Definition 2. Given QI table and STs with n attributes, the group id G_{ID_l}, the sensitive attribute SA, the breach probability of an individual when the adversary knows the target individual group id is as follows:

$$BP[t.[A_1, A_2, \ldots, A_n]|t.[A_{n+1}] = l] = \prod_{j=1}^{|SA|} \frac{|c_m^j|}{|G_{ID_l}|} = \frac{\prod_i^{|SA|} |c_m^j|}{|G_{ID_l}|^{|SA|}} \tag{3}$$

where $t.[]$ represents the tuple under all the attributes of the QI table and STs, $t.[A_{n+1}]$ represents the $(n+1)^{\text{th}}$ attribute instance and $|SA|$ is the number of sensitive attributes and $|c_m^j|$ is the count of the corresponding sensitive value.

Definition 3. Given QI table and STs having n attributes, the group id G_{ID_l}, the sensitive attribute SA, the breach probability of an individual when the adversary knows only one sensitive attribute value s_m^j, i.e. $t[SA_m] = s_m^j$ after reconstructing QI table and STs, is as follows:

$$BP\left[t.[A_1, A_2, \ldots, A_n]|t.[SA_m] = s_m^j\right] = \frac{|c_m^j|}{|G_{ID_l}|}. \tag{4}$$

Definition 4. Given QI table and STs having n attributes, the group id G_{ID_l} and the sensitive attribute SA, the 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, is given by:

$$\begin{aligned} BP\left[t.[A_1, A_2, \ldots, A_n]|t.[A_{n+1}] = l \& t.[SA_m] = s_m^j\right] &= \prod_{j=1 \& j \neq m}^{|SA|} \frac{|c_m^j|}{|G_{ID_l}|} \\ &= \frac{\prod_j^{|SA|} |c_m^j|}{|G_{ID_l}|^{|SA|-1}}. \end{aligned} \tag{5}$$

Based on the Definitions 2, 3 and 4 different lemmas and properties are derived. The definitions and detailed proofs of these are given in [32].

5.1 Break-Merge Algorithm

The algorithm Break-Merge is shown in the Algorithm 2. The algorithm has two phases: Assignment of group id and decoupling of the anonymized dataset D'. During the first phase the group id is assigned to D' (steps 2–9). First, each tuple in

Algorithm 2: Break-Merge

Input: An Anonymized Dataset ($D^{'}$)
Output: QI_{table} and S_{tables} (STs)
Assumptions:
$QI_{table} = \phi$
$S_{tables} = \{S_{table_1}, S_{table_2}, \ldots, S_{table_m}\} = \phi$,
$AT' = \phi$, $G_{cnt} = 1$;
$Q_{Set_i} = \{Q_{Set_1}, Q_{Set_2}, \ldots, Q_{Set_d}\}$;
$S_{Set_j} = \{S_{Set_1}, S_{Set_2}, \ldots, S_{Set_m}\}$;

1 **Begin**
2 **For each** Tuple T_i in D'
3 **If**($Q_{Set_i} == Q_{Set_{i+1}}$) **then**
4 $Insert_{tuple}$ $(Q_{Set_i}, G_{cnt}, S_{Set_i})$ into AT'
5 **Else**
6 $Insert_{tuple}$ $(Q_{Set_i}, G_{cnt}, S_{Set_i})$ into AT'
7 $G_{cnt} = G_{cnt} + 1$;
8 **End If**
9 **End For**

10 **For** i = 1 to G_{cnt}
11 **For each** tuple t_j in AT'_i
12 $Insert_{tuple}(Q_{Set_j}, i)$ into Q_{table};
13 **End For**
14 **End For**

15 **For each** S_k in $S_{Set}\{S_{set_1}, S_{set_2,\ldots,S_{set_m}}\}$
16 **For each** distinct S_k value v in AT'_i
17 $support_{i_j}$(v)=The number of records in ADT^*_i with sensitive value v;
18 $Insert_{tuple}(i, v, support_{i_j}(v))$ into S_{tables_j}
19 **End For**
20 **End For**
21 **End**

the Quasi group is compared with the $(n+1)^{\text{th}}$ tuple in the group and if they are equal, i.e. the same equivalence class, common group id is assigned. These steps are repeated for all the tuples in the anonymized data to form QI partitions with their corresponding group-ids. During the second phase the anonymized data is divided into Q-table and sensitive tables (STs). For each tuple t_j in AT'_i the tuples are inserted into QI_{table} which contains only the QI group associated with its corresponding group id having the form (Q_{Set_j}, i) (step 11) and when coming to the sensitive tables for each Quasi 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 QI_{table}, STs 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.

5.2 Limitations of Break-Merge

Break-Merge cannot identify breach probabilities when new tuples are added for the next release. In view of this following research questions could be raised.

- How many tuples must be added to the dataset such that privacy breach will not occur in the new release?
- Is there any threshold defined for the breach probability?
- Can we estimate the incremental ratio of the breach probability due to addition of new tuples?

The above research questions are dealt in the next section.

6 INCREMENTAL BREAK-MERGE

In this section, we extended the Break-Merge approach where the possibility of adding new tuples are ready for a new release of the dataset. We assume that the new dataset contains the same attributes as in the old release. In this paper, we do not consider the scenario of continuous data streams publishing (for example: credit card transactions) in which the timestamp of the new stream is considered and then anonymized accordingly.

Identifier	*Quasi-Identifier*			*Sensitive Attribute*		
Name	**Age**	**Sex**	**Zipcode**	**Government**	**Marital-Status**	**Salary**
...	...	...	...	...	...	...
Anne	42	F	17026	State-gov	Never-Married	≤ 50 K
Abby	39	F	13834	State-gov	Never-Married	≤ 50 K
Nancy	47	F	18002	State-gov	Never-Married	≤ 50 K
Mary	50	F	23046	State-gov	Never-Married	≤ 50 K
Mari	33	F	16523	State-gov	Never-Married	≤ 50 K
Lindy	34	F	22000	State-gov	Never-Married	> 50 K
Emma	41	F	23000	State-gov	Never-Married	> 50 K
Madison	32	F	16000	State-gov	Never-Married	> 50 K
Sarah	73	F	19000	State-gov	Never-Married	> 50 K
Grace	50	F	22000	State-gov	Never-Married	> 50 K
...	...	...	...	...	...	...

Table 9. $T_{Incremental}$ dataset

When new tuples are added to the original dataset as shown in the Table 9 the new dataset will be incremented. The whole dataset is re-anonymized accordingly.

Age	Sex	Zipcode	Government	Marital-Status	Salary
[30–50]	F	[13000–23000]	State-Gov	Never-Married	≤ 50 K
[30–50]	F	[13000–23000]	State-Gov	Never-Married	≤ 50 K
[30–50]	F	[13000–23000]	State-Gov	Never-Married	≤ 50 K
[30–50]	F	[13000–23000]	State-Gov	Never-Married	≤ 50 K
[30–50]	F	[13000–23000]	State-Gov	Never-Married	≤ 50 K
[30–50]	F	[13000–23000]	State-Gov	Never-Married	> 50 K
[30–50]	F	[13000–23000]	State-Gov	Never-Married	> 50 K
[30–50]	F	[13000–23000]	State-Gov	Never-Married	> 50 K
[30–50]	F	[13000–23000]	State-Gov	Never-Married	> 50 K
[30–50]	F	[13000–23000]	State-Gov	Never-Married	> 50 K

Table 10. $T_{incremental}$ anonymized tuples

Table 10 shows the newly anonymized tuples only. Due to space limit and for clear explanation we only considered two combinations, i.e., 5 females each who are working in state-gov and whose marital-status is never married and salary may be either $\leq$ 50 K or $>$ 50 K. Since we apply Break-Merge technique and break the dataset into QI and S tables the newly populated tables have new incremental counts of the corresponding sensitive values accordingly as shown in the Tables 11, 12 and 13.

Since the counts of the sensitive attributes have changed the likelihood has been re-calculated and shown in Table 14. It has been clearly observed that when the new tuples are added to the previous dataset the probabilities have been increased drastically. Due to space limit we tabulated few combinations and their likelihoods. This strengthens the adversary knowledge breach in identifying the individual. In order to limit the attackers breach probability the data publisher allows the addition of new tuples to the dataset until a desired risk level α is reached as defined earlier Property 3: The knowledge breach probability for the new incremental dataset $T_{Incremental}$ will be greater than the knowledge breach probability of a non-incremental dataset T. Formally it is defined below

$$BP_{inc} = \prod_{i=k+1}^{|SA|} \frac{|G_{ID_l}|XX[SA_i = v_i^j] + 1}{|G_{ID_l}| + 1}$$

$$> \prod_{i=k+1}^{|SA|} \frac{|G_{ID_l}|XX[SA_i = v_i^j]}{|G_{ID_l}|} > BP.$$

6.1 Incremental Break-Merge Algorithm

The algorithm for incremental Break-Merge algorithm is shown in Algorithm 3. This algorithm takes the new incremental dataset IncT(DS*), the quasi identifier table (QI) and the sensitive tables (ST's) that were generated by the Break-Merge

G_{ID}	**Marital-Status**	**Count**
1	Married	1
1	Never-Married	**12**
1	Divorced	1
1	Separated	1
2	Married	3
2	Never-Married	1
2	Divorced	1

Table 11. $T_{Incremental}$ sensitive table of Marital-Status

G_{ID}	**Salary**	**Count**
1	$\leq 50\,\mathrm{K}$	**9**
1	$> 50\,\mathrm{K}$	***6***
2	$\leq 50\,\mathrm{K}$	3
2	$> 50\,\mathrm{K}$	2

Table 12. $T_{Incremental}$ of Salary

algorithm as inputs. The algorithm updates the counts of the QIT and ST's as follows. Initially for each new tuple the breach probability is calculated (lines 2–4) and Break-Merge principle is applied. The new breach probability is verified for the defined breach probability limit α (line 5). Once this limit is reached no new tuples are further allowed. The number of records that are allowable is returned along with new incremental QI table and ST's.

6.2 Analysis of Incremental Break-Merge

As discussed earlier the Break-Merge technique provides an optimal solution for reducing the dependencies in an anonymized dataset. However, this technique has its own limitation as discussed in Section 5.3. In this section an analysis is done to show how the incremental breach probabilities vary while new tuples are added to the dataset. We presented different risk levels (α) by varying tuples.

G_{ID}	**Government**	**Count**
1	State-Gov	**12**
1	Federal-Gov	1
1	Private	1
1	Local-Gov	1
2	Private	3
2	Self-emp-not-inc	1
2	Federal-Gov	1

Table 13. $T_{Incremental}$ of Government

Algorithm 3: Incremental Break-Merge

Input: $Inc(D')$, QI_{table} and $S_{tables}(STs)$
Output: $IncQ_{table}$, $IncSTs$, $support(v)$
Assumptions:
$IncQ_{table} = \phi$,
$IncST = \{IncST_1, IncST_2, \ldots, IncST_m\} = \phi$,
$Q_{Set_i} = \{Q_{Set_1}, Q_{Set_2}, \ldots, Q_{Set_d}\}$;
$S_{Set_j} = \{S_{Set_1}, S_{Set_2}, \ldots, S_{Set_m}\}$;

```
1  Begin
2    For each Tuple T_i in Inc(D')
3      kbp = 1
4      For j = 1 to |IncST|
5      kbp = kbp * count_j / G_cnt_j
6    If kbp ≤ α  then
7        Call Break-Merge
7    Else
9        return support(v)
         // The Permitted number of
            records with sensitive value v
10 End
```

Different alpha values, let us say 0.6, 0.7, 0.87, 0.9, are considered for different tuples ranging from 10 to 200. It can be easily observed that the data publisher can assess for the defined risk level of the publisher and can determine for which number of tuples the risk level exceeds. For instance when you see Figure 15 a) (Section 7, Phase III) the risk level is set to 0.6 and maximum number of tuples that can be added to the dataset are 17. If the number of tuples exceeds the breach probability increases drastically. We varied the number of tuples from 10 to 200 for different risk levels. When more than 190 tuples are added the risk level reaches 0.9 which shows a very high vulnerable breach Figure 15 d). This technique leverages the data publisher for making decisions on the release of the dataset. We also showed how much was the probability increased when different combinations of new tuples were added to the older release. Table 14 shows Incremental ratio for various combinations of dependencies. For instance when you see the dependency combination State-gov $\rightarrow$ NeverMarried (NM) and Salary $>$ 50 K the breach probability increased 3 times when 19 new tuples are added to the new release. This identification will help the data publisher to determine the number of tuples that must be added and the addition of tuples can be restricted until the desired threshold α is reached. Figures 11, 12, 13 and 14 show the probabilities increase for various different combinations.

Gov	**MS**	**Salary**	**Prob**	**Like**	$Prob_{Inc}$	$Like_{Inc}$	Inc_{Ratio}
State-Gov	NM	≤ 50 K	$\frac{2}{5} * \frac{4}{5}$	32 %	$\frac{19}{23} * \frac{14}{23}$	50 %	0.5625
State-Gov	NM	> 50 K	$\frac{2}{5} * \frac{1}{5}$	8 %	$\frac{19}{23} * \frac{9}{23}$	32 %	3
State-Gov	NM	≤ 50 K	$\frac{2}{5} * \frac{4}{5}$	32 %	$\frac{21}{25} * \frac{15}{25}$	50 %	0.5625
State-Gov	NM	> 50 K	$\frac{2}{5} * \frac{4}{5}$	8 %	$\frac{21}{25} * \frac{10}{25}$	34 %	3.25
State-Gov	NM	≤ 50 K	$\frac{2}{5} * \frac{4}{5}$	32 %	$\frac{30}{35} * \frac{21}{35}$	51 %	0.593
State-Gov	NM	> 50 K	$\frac{2}{5} * \frac{4}{5}$	8 %	$\frac{30}{35} * \frac{21}{35}$	34 %	3.25
State-Gov	NM	≤ 50 K	$\frac{3}{5} * \frac{4}{5}$	32 %	$\frac{39}{45} * \frac{27}{45}$	52 %	0.625
State-Gov	NM	> 50 K	$\frac{3}{5} * \frac{4}{5}$	8 %	$\frac{39}{45} * \frac{18}{45}$	35 %	3.375
Federal	Married	≤ 50 K	$\frac{1}{5} * \frac{4}{5}$	16 %	$\frac{3}{7} * \frac{6}{7}$	37 %	1.3125
Federal	Married	> 50 K	$\frac{1}{5} * \frac{1}{5}$	4 %	$\frac{3}{7} * \frac{1}{7}$	6 %	0.5
Federal	Married	≤ 50 K	$\frac{1}{5} * \frac{4}{5}$	16 %	$\frac{7}{13} * \frac{10}{13}$	41 %	1.562
Federal	Married	> 50 K	$\frac{1}{5} * \frac{1}{5}$	4 %	$\frac{7}{13} * \frac{3}{13}$	12 %	2.5
Federal	Married	≤ 50 K	$\frac{1}{5} * \frac{4}{5}$	16 %	$\frac{11}{20} * \frac{15}{20}$	41 %	2.43
Federal	Married	> 50 K	$\frac{1}{5} * \frac{1}{5}$	4 %	$\frac{11}{20} * \frac{5}{20}$	14 %	2.75
Federal	Married	≤ 50 K	$\frac{1}{5} * \frac{4}{5}$	16 %	$\frac{26}{37} * \frac{29}{37}$	55 %	0.625
Federal	Married	> 50 K	$\frac{1}{5} * \frac{1}{5}$	4 %	$\frac{26}{37} * \frac{8}{37}$	15 %	0.75
Private	Divorced	≤ 50 K	$\frac{1}{5} * \frac{4}{5}$	16 %	$\frac{3}{9} * \frac{7}{9}$	26 %	1.25
Private	Divorced	> 50 K	$\frac{1}{5} * \frac{1}{5}$	4 %	$\frac{3}{9} * \frac{2}{9}$	7 %	1.75
Private	Divorced	≤ 50 K	$\frac{1}{5} * \frac{4}{5}$	16 %	$\frac{6}{13} * \frac{10}{13}$	36 %	1.25
Private	Divorced	> 50 K	$\frac{1}{5} * \frac{1}{5}$	4 %	$\frac{6}{13} * \frac{3}{13}$	11 %	1.75
Private	Divorced	≤ 50 K	$\frac{1}{5} * \frac{4}{5}$	16 %	$\frac{14}{24} * \frac{17}{24}$	41 %	1.56
Private	Divorced	> 50 K	$\frac{1}{5} * \frac{1}{5}$	4 %	$\frac{14}{24} * \frac{7}{24}$	17 %	3.25
Private	Divorced	≤ 50 K	$\frac{1}{5} * \frac{4}{5}$	16 %	$\frac{22}{37} * \frac{26}{37}$	42 %	1.625
Private	Divorced	> 50 K	$\frac{1}{5} * \frac{1}{5}$	4 %	$\frac{22}{37} * \frac{11}{37}$	18 %	3.5

Table 14. Knowledge breach probability

7 EXPERIMENTATION

Experimentations are conducted in three 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 [41]. In Phase III we show how the knowledge breach probabilities increase with the increment of new records for different risk values defined by the data publisher as explained in Sections 6.1 and 6.2.

7.1 Phase I

Experiments were conducted on Adult dataset available at UCI machine Learning Repository [39]. 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, sex and zipcode were treated as quasi-identifiers and government, marital-status and salary were treated as sensitive attributes. Weka tool was used to

construct the Bayesian net [6]. Experiments were conducted on both single sensitive attributes and multiple sensitive attributes. The dataset is replicated such that each equivalence size is 1 000. For the construction of Bayesian network it took less than 1.2 secs and nearly 1.7 secs for single and multiple sensitive attributes respectively (Figures 5 and 6) for a dataset with 100 000 records.

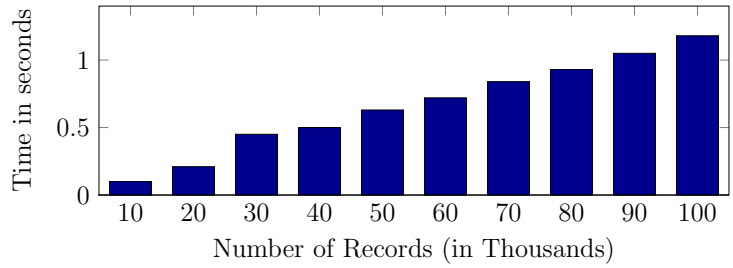

Figure 5. Time taken for constructing Bayesian net for dataset with single sensitive attribute

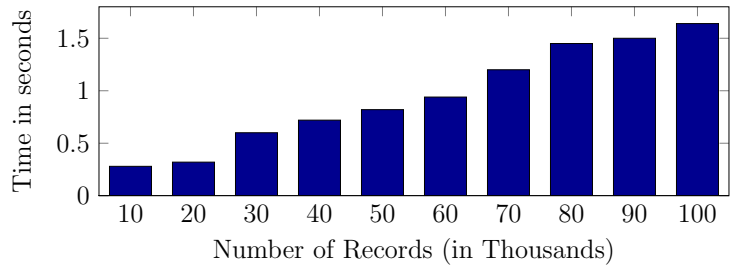

Figure 6. Time taken for constructing Bayesian net for dataset with three sensitive attributes

7.2 Phase II

The experimental setup for Break-Merge technique is the 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 [41]. They construct 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 breaks the table into QI and Sensitive tables as explained in Section 4.

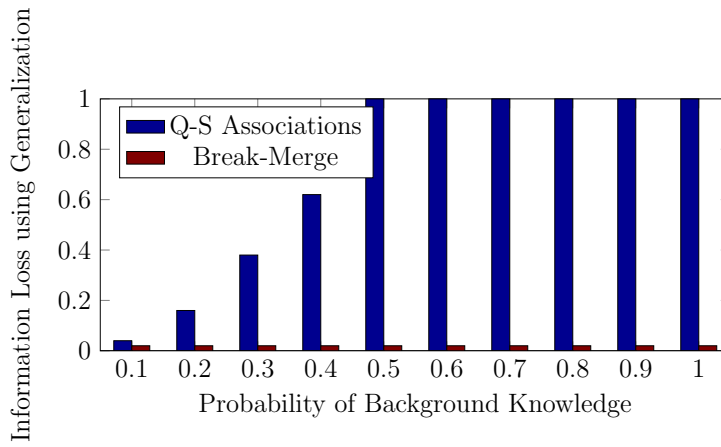

Figure 7. Comparison between Q-S association and BM

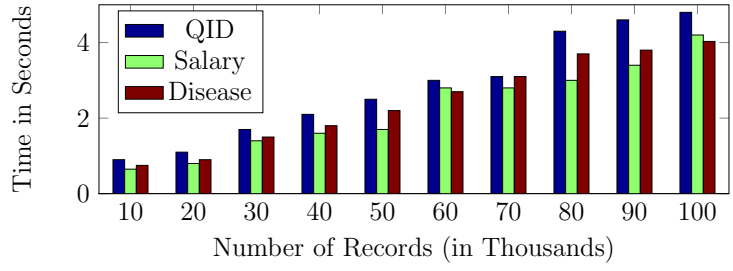

Figure 8. Breaking time for Adult dataset

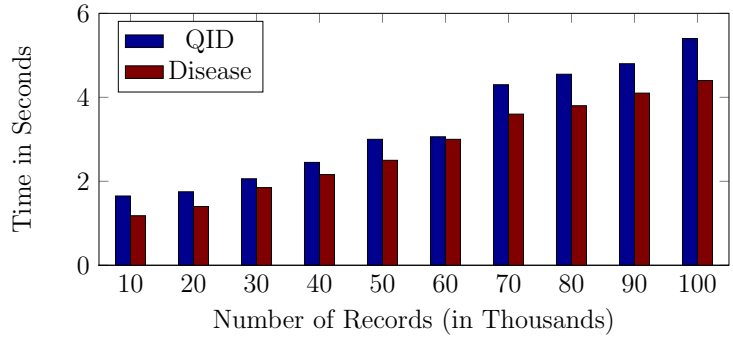

Figure 9. Breaking time for Disease dataset

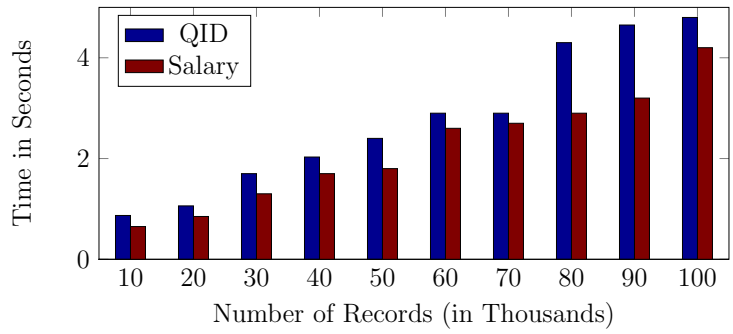

Figure 10. Breaking time for Salary dataset

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 thereafter remain in that high level in Q-S association, but our approach does not generalize the sensitive attributes and the information loss with respect to the sensitive attributes will be zero. Different performance measures for breaking the tables into 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 Adult dataset that contains 100 000 records, as shown in Figure 8. The remaining datasets Disease and Salary also took less than 5 seconds, as shown in Figures 9 and 10, respectively.

7.3 Phase III

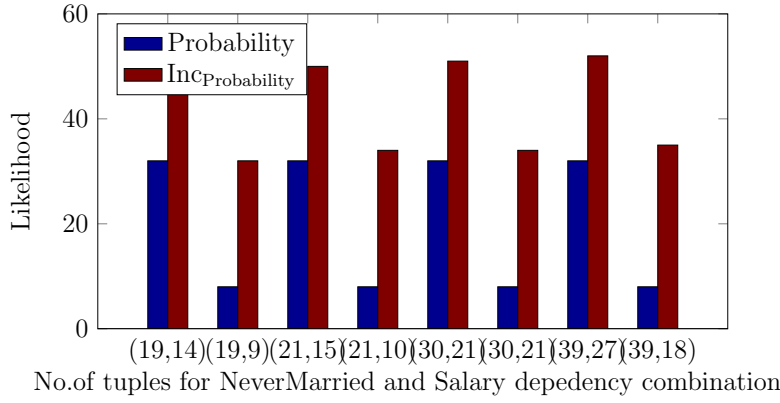

Figure 11. Comparison between Break-Merge and Incremental Break-Merge breach probabilities

[H]

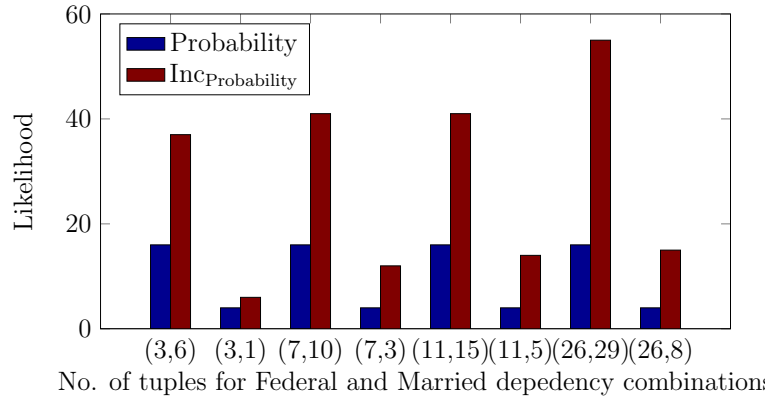

Figure 12. Comparison between Break-Merge and Incremental Break-Merge breach probabilities

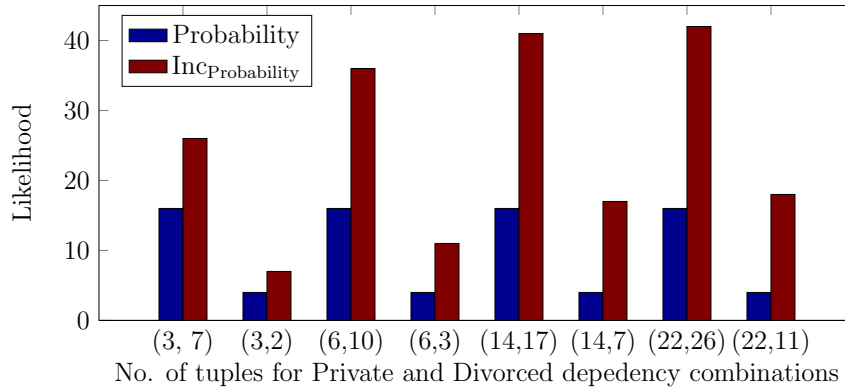

Figure 13. Comparison between Break-Merge and Incremental Break-Merge breach probabilities

8 CONCLUSIONS AND FUTURE DIRECTIONS

In this current internet era especially in information dissemination sectors, Privacy Preserving Data Publishing (PPDP) has become an interesting and challenging problem to deal with. Much of the research focused on anonymizing the data using different frameworks and principles. But, the focus on verifying the data once it is published is very trivial. The publisher is releasing the anonymized data without any prior verification of how the adversary could identify an individual with

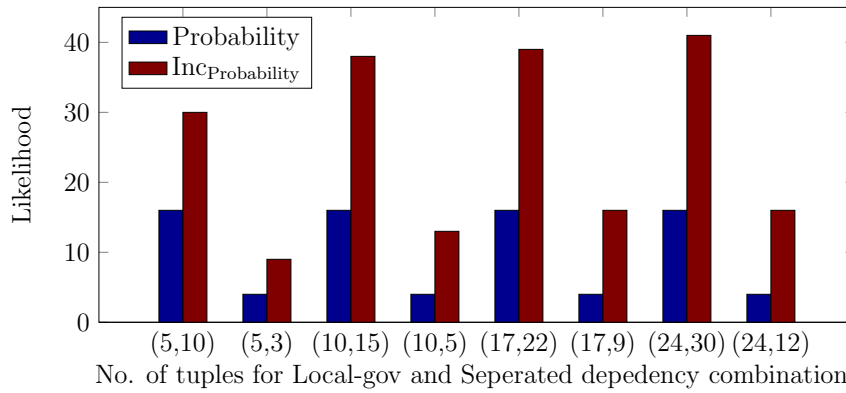

Figure 14. Comparison between Break-Merge and Incremental Break-Merge breach probabilities

his background knowledge and thereby compromising the privacy. In this paper, we presented and discussed different folds of verification techniques on the published data. We unfolded and contributed various solutions to achieve the following objectives:

- To detect dependencies among quasi-identifiers and Sensitive attributes and determine the breach probability
- To reduce the identified dependencies
- To identify breach probability for every incremental release and
- To determine the limit on the addition of new data tuples for every new release.

Our first contribution: we identified dependencies in an anonymized dataset by constructing belief networks on different possibilities of the attributes. Our analysis showed how many breach possibilities can be caused by an attacker when he/she possesses hypothetical information of an individual. We further showed how the data publisher can understand the nature of those vulnerable records by observing the breach probability based on the defined threshold.

Secondly, we presented how these dependencies can be reduced by our proposed verification model. In our model, the Break-Merge approach reduces the knowledge breach. It simply decouples the anonymized dataset into Quasi table and Sensitive tables. The main principle it follows is that no direct link exists between the generalized group and the sensitive values. Here the sensitive values are not generalized or suppressed as done in Q-S associations technique. This enhances our proposed technique by showing how the breach probability was increasing drastically for different risk levels and it also gives a better utility during data mining.

Finally, when subsequent release of data is considered we showed how *Break-Merge* limits in not analyzing the new data that may reveal the dataset to attack.

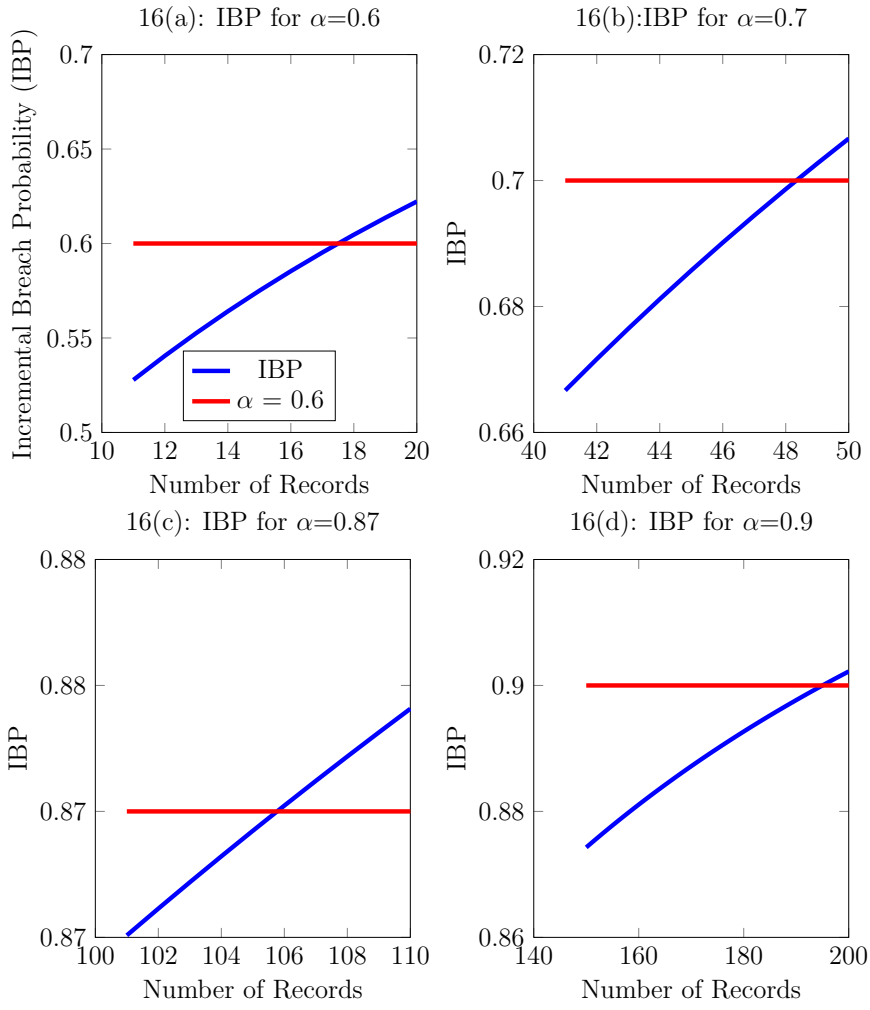

Figure 15. Incremental knowledge breach probability

Our new technique *Incremental Break-Merge* is a refined Break-Merge technique where the verification is done by adding each tuple and then repeatedly checked until the desired threshold is arrived. This signifies to what level and what amount of data must be added to reduce the risk levels considerably. When once the desired threshold is reached we stop adding new records. A comparison is made between Break-Merge and Incremental Break-Merge to show how the likelihood increases reasonably high when adding new records. To our knowledge, our verification techniques are first of the kind in the literature. We limited our study only for different static releases of data. In future we need to further investi-

gate how the nature of the likelihood may change in the dynamic release of the data.

REFERENCES

[1] Adam, N. R.—Wortmann, J. C.: Security-Control Methods for Statistical Databases: A Comparative Study. ACM Computing Survey, Vol. 21, 1989, No. 4, pp. 515–556.

[2] Aggarwal, C. C.: On k-Anonymity and the Curse of Dimensionality. In: Böhm, K. et al. (Eds.): Proceedings of the 31st International Conference (VLDB '05), Trondheim, Norway, 2005, pp. 901–909.

[3] Barak, B.—Chaudhuri, K.—Dwork, C.—Kale, S.—McSherry, F.—Talwar, K.: Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release. Proceedings of the SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS '07), Beijing, China, June 11–14, 2007, pp. 273–282.

[4] Bayardo, R.—Agrawal, R.: Data Privacy Through Optimal k-Anonymization. Proceedings of the International Conference on Data Engineering (ICDE '05), Tokyo, Japan, April 5–8, 2005, pp. 217–228.

[5] Fung, B. C. M.—Wang, K.—Fu, A. W.-C.—Yu, P. S.: Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques. CRC Press, 2011.

[6] Bouckaert, R. R.: Bayesian Network Classifiers in Weka for Version 3-5-5. The University of Waikato, 2007.

[7] Burnett, L.—Barlow-Stewart, K.—Proos, A.—Aizenberg, H.: The "Gene Trustee": A Universal Identification System That Ensures Privacy and Confidentiality for Human Genetic Databases. Journal of Law and Medicine, Vol. 10, 2003, No. 4, pp. 506–513.

[8] Dalvi, N.—Miklau, G.—Suciu, D.: Asymptotic Conditional Probabilities for Conjunctive Query. Proceedings of the 10th Internationl Conference on Database Theory (ICDT '05). Springer, LNCS, Vol. 3363, 2005, pp. 289–305.

[9] Deustch, A.—Papakonstantinou, Y.: Privacy in Database Publishing. Proceedings of the 10th International Conference on Database Theory (ICDT '05). Springer, LNCS, Vol. 3363, 2005, pp. 230–245.

[10] Dwork, C.: Differential Privacy. Proceedings of International Colloquium on Automata, Languages and Programming (ICALP 2006). Springer, LNCS, Vol. 4052, 2006, pp. 1–12.

[11] Dwork, C.—McSherry, F.—Nissim, K.—Smith, A.: Calibrating Noise to Sensitivity in Private Data Analysis. Proceedings of the Third Conference on Theory of Cryptography (TOC '06). Springer, LNCS, Vol. 3876, 2006, pp. 265–284.

[12] Fung, B. C. M.—Wang, K.—Yu, P. S.: Top-Down Specialization for Information and Privacy Preservation. Proceedings of the International Conference on Data Engineering (ICDE '05). IEEE, Tokyo, Japan, April 5–8, 2005, pp. 205–216.

[13] Ghinita, G.—Karras, P.—Kalnis, P.—Mamoulis, N.: Fast Data Anonymization with Low Information Loss. Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB '07), September 23–27, 2007, Vienna, Austria, pp. 758–769.

[14] Hundepool, A.—Willenborg, L.: μ-Argus: Software for Statistical Disclosure Control. International Seminar on Statistical Confidentiality, 1996.

[15] Iyengar, V.: Transforming Data to Satisfy Privacy Constraints. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Edmonton, Alberta, Canada, July 23–26, 2002, pp. 279–288.

[16] Li, J.—Wong, R. C.-W.—Fu, A. W.-C.—Pei, J.: Anonymization by Local Recoding in Data with Attribute Hierarchical Taxonomies. IEEE Transactions on Knowledge and Data Engineering, Vol. 20, 2008, No. 9, pp. 1181–1193.

[17] Kim, J.: Method for Limiting Disclosure of Microdata Based on Random Noise and Transformation. Survey Research Methods of the American Statistical Association, 2001, pp. 328–387.

[18] LeFevre, K.—DeWitt, D. J.—Ramakrishnan, R.: Mondrian Multidimensional k-Anonymity. Proceedings of the 22nd International Conference on Data Engineering (ICDE '06), IEEE Computer Society, April 3–8, 2006, Atlanta, GA, USA, pp. 25.

[19] LeFevre, K.—DeWitt, D. J.—Ramakrishnan, R.: Workload-Aware Anonymization. Proceedings of the Twelfth Annual SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20–23, 2006, Philadelphia, USA, pp. 277–286.

[20] Li, N.—Li, T.—Venkatasubramanian, S.: t-Closeness. Privacy Beyond k-Anonymity and l-Diversity. Proceedings of the 23rd International Conference on Data Engineering, IEEE, April 15–20, 2007, Istanbul, Turkey, pp. 106–115.

[21] Machanavajjhala, A.—Gehrke, J.: On the Efficiency of Checking Perfect Privacy. Proceedings of the Twenty-Fifth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, ACM, June 26–28, Chicago, Illinois, USA, pp. 163–172.

[22] Machanavajjhala, A.—Kifer, D.—Gehrke, J.—Venkitasubramaniam, M.: l-Diversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discovery from Data, Vol. 1, 2007, No. 1, Art. No. 3..

[23] Martin, D. J.—Kifer, D.—Machanavajjhala, A.—Gehrke, J.—Halpern, J. Y.: Worst-Case Background Knowledge for Privacy Preserving Data Publishing. Proceedings of the 23rd International Conference on Data Engineering, IEEE, April 15–20, Istanbul, Turkey, pp. 126–135.

[24] Meyerson, A.—Williams, R.: On the Complexity of Optimal k-Anonymity. Proceedings of the Twenty-Third ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, ACM, June 14–16, 2004, Paris, France, pp. 223–228.

[25] Miklau, G.—Suciu, D.: A Formal Analysis of Information Disclosure in Data Exchange. Proceedings of the ACM SIGMOD/PODS Conference, ACM, June 13–18, 2004, Paris, France, pp. 575–586.

[26] Neapolitan, R. E.: Learning Bayesian Networks. Pearson Education, 2004.

[27] Nissim, K.—Raskhodnikova, S.—Smith, A.: Smooth Sensitivity and Sampling in Private Data Analysis. Proceedings of the 39th ACM Symposium on Theory of Computing, ACM, June 11–13, 2007, San Diego, California, USA, pp. 75–84.

[28] Pearl: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, San Franciso, CA, USA, 1998.

[29] Samarati, P.—Sweeney, L.: Generalizing Data to Provide Anonymity When Disclosing Information. Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, ACM, June 1–3, 1998, Seattle, Washington, USA, pp. 188.

[30] Samarati, P.: Protecting Respondents' Identities in Microdata Release. Proceedings of the IEEE Transcations on Knowledge and Data Engineering, Vol. 13, 2001, pp. 1010–1027.

[31] Sandeep Varma, N.—Valli Kumari, V.: BM (Break-Merge): An Elegant Approach for Privacy Preserving Data Publishing. Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk and Trust, and IEEE International Conference on Social Computing, October 9–11, 2011, MIT, Boston, USA, pp. 1202–1207.

[32] Sandeep Varma, N.—Valli Kumari, V.: Detecting Dependencies in an Anonymized Dataset. Proceedings of the International Conference on Advances in Computing, Communications and Informatics, ACM, August 3–5, Chennai, pp. 82–89.

[33] SAS Institute: SAS/STAT 9.2 User's Guide. 1st Edition. SAS Publishing, 2008.

[34] SPSS Inc.: SPSS 16.0 Base User's Guide. 2nd Edition. SPSS Inc., 2007.

[35] Stata Corporation: Stata User's Guide Release 8.0. 1st Edition. Stata Press, 2003.

[36] Sweeney, L.: Datafly: A System for Providing Anonymity in Medical Data. Proceedings of the Eleventh International Conference on Database Security, August 10–13, 1997, California, USA, pp. 356–381.

[37] Sweeney, L.: k-Anonymity – A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge Based Systems, Vol. 10, 2002, No. 5, pp. 557–570.

[38] Sweeney, L.: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. International Journal on Uncertainty, Fuzziness and Knowledge Based Systems, Vol. 10, 2002, No. 5, pp. 571–588.

[39] UCI Web Site. Available on `http://www.ics.uci.edu`.

[40] Verma, T.—Pearl, J.: An Algorithm for Deciding if a Set of Observed Independencies Has a Causal Explanation. Proceedings of the Eighth Annual Conference on Uncertainty in Artificial Intelligence, July 17–19, 1992, Stanford University, Stanford, CA, USA, pp. 323–330.

[41] Weijia, Y.—Shangten, H.: k-Anonymization without Q-S Associations. Proceedings of the Joint 9th Asia-Pacific Web and 8th International Conference on Web-Age Information Management Conference on Advances in Data and Web Management. Springer, LNCS, Vol. 4505,, 2007, pp. 753–764.

[42] Wendy, H. W.: Ambiguity: Hide the Presence of Individuals and Their Privacy with Low Information Loss. Proceedings of the International Conference on Management of Data, CSI, December 17–19, 2008, Mumbai, India.

[43] Wong, R. C. W.—Li, J.—Fu, A. W.-C.—Wang, K.: (α, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing. Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20–23, 2006, Philadelphia, USA, pp. 754–759.

[44] Xiao, X.—Tao, Y.: Personalized Privacy Preservation. Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, June 27–29, 2006, Chicago, Illinois, USA, pp. 229–240.

[45] Xiao, X.—Tao, Y.: Anatomy: Simple and Effective Privacy Preservation. Proceedings of the 32nd International Conference on Very Large Data Bases, September 12–15, 2006, Seoul, Korea, pp. 139–150.

[46] Xiao, X.—Tao, Y.: m-Invariance: Towards Privacy Preserving Re-Publication of Dynamic Datasets. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 11–14, 2007, Beijing, China, pp. 689–700.

[47] Xu, J.—Wang, W.—Pei, J.—Wang, X.—Shi, B.—Fu, A. W.-C.: Utility Based Anonymization Using Local Recoding. Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20–23, 2006, Philadelphia, USA, pp. 785–790.

[48] Zhang, Q.—Koudas, N.—Srivastava, D.—Yu, T.: Aggregate Query Answering on Anonymized Tables. Proceedings of the 23rd International Conference on Data Engineering, IEEE, April 15–20, 2007, Istanbul, Turkey, pp. 116–125.

Sandeep Varma Nadimpalli is currently working as Assistant Professor in Department of Information Science and Engineering, BMS College of Engineering. He received his B.Tech. degree in information technology from JNTU Hyderabad, Telangana, India in 2007. He received his M.Tech. from Andhra University in 2009 and his Ph.D. in computer science and systems engineering from Andhra University in 2015. He also worked as Junior Research Fellow (Professional) from 2009 to 2011 and later worked as Senior Research Fellow from 2011 to 2014 at Andhra University. His research interests include data privacy, formal verifi cation techniques on security protocols, software metrics and cloud computing. He is a member of IEEE.

Valli Kumari Vatsavayi is currently Professor in Computer Science and Systems Engineering Department and is also Honorary Director to 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 she has 100 publications in various conferences and journals of international and national repute. She is an active consultant to several government, public sector, private organisations. She is an active member of IEEE, ACM, CRSI and CSI. She also holds a prestigious position as Vice-Chair for IEEE Vizag Bay Subsection.