# Incremental Diversity: An Efficient Anonymization Technique for PPDP of Multiple Sensitive Attributes

Veena Gadad, Sowmyarani C N
Department of Computer Science and Engineering
R V College of Engineering, Karnataka, India

***Abstract*—Data collected at the organizations such as schools, offices, healthcare centers and e-commerce websites contain multiple sensitive attributes. The sensitive information from these organisations such as marks obtained, salary, disease, treatment and traveling history are personal information that an individual dislikes to disclose to the public as it may lead to privacy threats. Therefore, it is necessary to preserve privacy of the data before publishing. Privacy Preserving Data Publishing(PPDP) algorithms aim to publish the data without compromising the privacy of individuals. In the recent years several algorithms have been designed for PPDP multiple sensitive attributes. The major limitations are, firstly among several sensitive attributes these algorithms consider one of them as primary sensitive attribute and anonymize the data, however there may be other dominant sensitive attributes that need to be preserved. Secondly, there is no consistent way to categorize multiple sensitive attributes. Lastly, increased proportion of records are generated due to usage of generalization and suppression techniques. Hence, to overcome these limitations the current work proposes an efficient approach to categorize the sensitive attributes based their semantics and anonymize the data using an anatomy technique. This reduces the residual records as well as categorizes the attributes. The results are compared with popular techniques like Simple Distribution of Sensitive Values (SDSV) and (l, e) diversity. Experiments prove that our method outperforms the existing methods in terms of categorization of multiple sensitive attributes, reducing the percentage of residual records and preventing the existing privacy threats.**

***Keywords*—Data management; privacy preserving data publishing; data privacy; multiple sensitive attributes; data anonymization; privacy attacks***

## I. Introduction

The developments in digital devices and information systems have created various opportunities and challenges. Enormous amount of data gets collected by various digital devices in sectors such as healthcare, education, e-commerce, banking, government etc., and stored in the information systems. The data that is specific to a single organization is called as Microdata, among other attributes it contains individual's sensitive attributes.The main purpose of data collection is to glean actionable insights and help the organizations to perform analysis, research and succeed in terms of greater productivity and return on investments.Few organizations like healthcare centres, education and e-commerce share the microdata to third parties for investigation or stored in cloud and made available for researchers to perform some fact-findings [1].

Amid constructive usage of the microdata, there may be an intruder, the purpose is to steal individual information and cause privacy threats. Fig. 1 shows the process of micro data collection, storage, publishing and usage. The primary data is one that is collected directly from the source and contains personal information such as marks obtained, salary, credit card information, treatment history and disease. When such a data is shared to the public care must be taken not to disclose individuals sensitive information. Privacy Preserving Data Publishing (PPDP) provides methods and tools with the aim to protect the privacy of the individuals and at the same time make sure that is the data is usable by the public for analysis [2].

Anonymization algorithms are approaches that are commonly used to achieve PPDP [2] [3]. Existing algorithms were designed over a duration to overcome various privacy threats [4]. These algorithms can be broadly classified into algorithms that preserve privacy of Single Sensitive Attribute(PPDP-SSA) and those that preserve for Multiple Sensitive Attributes (PPDP-MSA).

1. PPDP-SSA: K- Anonymity [5]–[7] was the first anonymization model, it failed to prevent homogeneity and background knowledge attack [8]. Hence, l-diversity [9],t-closeness [10], Anatomy [11], Slicing [12] permutation based [13]–[15] algorithms were designed. Although, these algorithms surpassed the previous ones, the semantic relationship between the attributes was not considered. This resulted into new attacks, namely, similarity and semantic privacy attacks [16], [17] which were addressed in the next set of algorithms [17] [29] [30].

2. PPDP-MSA: With big data, IOT and cloud storage the microdata in effect consist of MSA that had to be preserved [18]–[20]. Many algorithms are proposed under this category [21]–[25], but the major limitations of these algorithms were: i) one of the attributes is chosen as a primary sensitive attribute and the data is anonymized, the other dominant sensitive attributes are not preserved. ii) The algorithms do not provide any basis for categorizing the sensitive attributes. iii) The algorithms use generalization and suppression techniques to anonymize the data which leads to generation of residual records.

Simple Distribution of Sensitive Values (SDSV) [26] is the recent approach that discusses distribution of MSA. Here, the ranking of the sensitive attributes is based on the frequency of occurrence. The author uses l-diversity to group the records. However, the approach do not consider the semantic similarity between the attributes, hence the data anonymized using this approach is vulnerable to semantic attacks.

In the current work, the semantic hierarchical trees are constructed for sensitive attributes of the microdata, based on

TABLE I. SAMPLE MICRODATA OF A HEALTH CARE CENTRE

| Name | Age | Gender | ZipCode | Disease | MartialStatus | Salary |
|---|---|---|---|---|---|---|
| Alice | 23 | F | 560098 | Flu | Unmarried | 45K |
| Bob | 25 | M | 560096 | Pneumonia | Married | 48K |
| Trudy | 30 | M | 560190 | Flu | Unmarried | 30K |
| Sophi | 36 | F | 560091 | Bronchitis | Unmarried | 55K |
| Tom | 39 | M | 560094 | HeartInfection | Married | 57K |
| Ellen | 42 | F | 560099 | HeartAttack | Married | 65K |
| Jessi | 52 | F | 560298 | GastricUlcer | Married | 45K |
| Paul | 53 | M | 560090 | Dyspepsia | Unmarried | 45K |
| Steve | 61 | M | 560092 | HeartInfection | Unmarried | 40K |

the similarity indicator 'e' proposed in (l, e)-diversity [17], the sensitive attributes are categorized into primary, secondary, tertiary sensitive attributes. Later, the records of the microdata are recursively grouped into the equivalence class such that each class satisfies l-diversity [9]. The results obtained after conducting the experiments prove that the proposed algorithm is efficient in terms of preventing the existing privacy threats associated with MSA, reducing the generation of residual records and providing a basis for categorizing the sensitive attributes.

### A. Organization of the Paper

The paper is organized as follows: Section II presents Data anonymization and Basic definitions, Section III presents the related work, Section IV discusses the proposed method and empirical results, Experiments and Performance Equations are presented in Section V. Results and Discussion are presented in Section VI. Section VI discusses Conclusion and Future work.

## II. DATA ANONYMIZATION AND BASIC DEFINITIONS

Data anonymization is a process of protecting individual's sensitive information so as to prevent disclosures and privacy threats. Fig. 1 shows the process of micro data collection, storage, publishing and usage. The collected data contains Multiple Sensitive Attributes (MSA) such as disease, treatment, salary, marks obtained, travel history and health conditions. The data owners dislike such data to be disclosed to others.

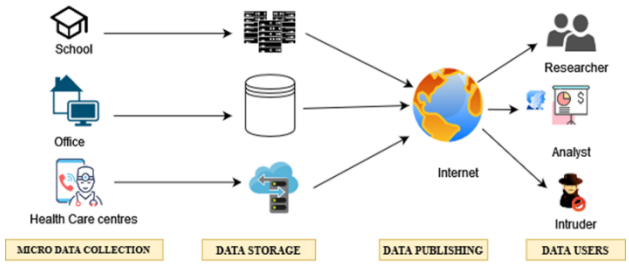


Fig. 1. Process of microdata collection, storage, publishing and usage.

Consider Table I, that is a sample microdata of a health care data center(M). From literature, the microdata attributes are classified as identifiers, quasi identifiers, sensitive and non sensitive attributes.

These attributes are defined as follows:

Identifiers (ID) – Directly identifying attributes are called as identifiers. For example: Name, Patient ID, Social Security number etc. Such attributes are removed be-fore publishing the micro data.

Quasi Identifiers (QID) – These attributes are used to indirectly identify a particular person. For example: Age, Gender and ZipCode. In any anonymization algorithms, the QID's are treated to different values to prevent disclosures.

Sensitive Attributes (SA) – The attributes that provide valuable information to the researchers/analyst and are used in data analysis. For example: Disease, Salary and Marital Status.

The following are the definitions of the terms that are used throughout the paper.

Definition 1: Equivalence class – Let a microdata M =(A1, A2, A3.. An) be the collection of records. The attributes are combinations of QID's and SA. 'n' is number of attributes. The equivalence class( EQ) is a group of records with indistinguishable mapping of QID values.

Definition 2: (e-Similar) - Let a1 and a2 be the levels of two sensitive values v1,v2 in their semantic tree respectively. A0 be the closest common ancestor. e= [(a1-a0)+(a2-a0)]/2. v1 and v2 are now said to be e-similar. In other words, the similarity between v1 and v2 is 'e'.

Definition 3: (l, e) Diversity - A data set is said to satisfy (l,e ) diversity[14] if every EQ is l-diverse and the similarity among any two values in an EQ is equal to or more than 'e'.

Definition 4: Anatomy – An Anatomy [8] is anonymization technique, it disassociates the sensitive at-tributes and quasi identifier attributes into two tables. These tables increases utility when compared to k-anonymity because the attribute values are published in its original form. The anatomy breaks the correlation between the SA and QID's, this increases the privacy.

Definition 5: Residue Records - Those records that do not fit into any equivalence classes as they do not satisfy the constraints of the equivalence class are called as residues. When any anonymization algorithms is applied care must be taken to ensure that the residue percentage is as less as possible.

## III. RELATED WORK

The datasets used in technologies such as BigData, IOT and cloud computing contain multiple sensitive attributes that need to be preserved [18]. Fig. 2 shows the advancement of the anonymization algorithms.

Initially, the algorithms were designed for SSA for example: k- anonymity [5], [7], [27], l-diversity [9], t-closensess [10], anatomy [11],slicing [12], failed to consider semantics between the sensitive atributes. Similarity and semantic similarity attacks [17] occur when the anonymization algorithms do not consider the semantics between the sensitive attributes. For example, Gastritis, Gastric Ulcer and Gastroparesis are diseases related to stomach. An intruder who has some background knowledge about the person can get to know that he is suffering from stomach infection.

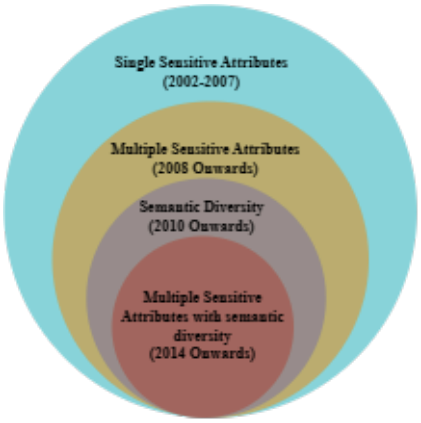


Fig. 2. Advancement of anonymization algorithms based on number of sensitive attributes and their semantics.

Later, algorithms like p-sensitive k-anonymity [28], (p+, $\alpha$) sensitive k-anonymity [29], (p+, $\alpha$, t) anonymity [30] were proposed. These algorithms,though considered the semantic relationship between the attributes, failed miserably when applied for dataset with MSA. Therefore, new set of algorithms were proposed to protect MSA.

The algorithms such as Rating [31], p-cover k- anonymity [32],Decomposition [33], Decomposition+ [34], KC slice [35], KCi slice [36] were designed to prevent privacy threats that occured on data with multiple sensitive attributes such as association attacks [37], semantic similarity attacks [16].In these algorithms, one of the attributes was considered as a primary sensitive attribute and other as secondary attribute. l-diversity, Anatomy or Slicing methods were used to group the records and anonymize the data. These algorithms did not discuss any method on how to select the sensitive attributes.

Simple Distribution of Sensitive Values for MSA (SDSV) [26] is a recent approach to distribute the MSA. In this method, two sensitivity levels are considered- High Sensitive Value (HSV) and Low Sensitive Value (LSV). Those sensitive attributes that have more than HSV is considered to be Primary Sensitive Attribute (PSA) and others are called as Contributory Sensitive Attributes (CSA). To understand this approach let us see basic definitions.

### *A. SDSV Approach to Select and Distribute the Multiple Sensitive Attributes*

The user first selects the High Sensitive Values(HSV's) that he wants to preserve. Consider Table I, let the selected HSV be 'Heart Infection", "$> 50K$" and 'Unmarried' for Disease, Salary and Marital Status attributes respectively. The occurrences of these values is II, III and IV in the table. Since the attribute value 'Unmarried' occurrence is high, the attribute Marital Status is considered as Primary Sensitive Attribute (PSA) and Disease, Salary are treated as Contributory Sensitive At-tribute(CSA). The table is anonymized as per anatomy [9]and it is published. The resulting tables are Table II, Table III , Table IV and Table V. Table II contains the QID's of the Microdata, these are grouped and assigned the GroupID. Table III contains the Marital Status as PSA that is grouped such that within each EQ there is equal diversity of the attribute value. Similarly, Table IV and Table V contains the grouping for Salary and Disease attributes. The groupID is assigned for the EQ's that are created in the tables.

TABLE II. QID TABLE

| Age | Gender | Zip code | GroupID |
|---|---|---|---|
| 23 | F | 560098 | 1 |
| 25 | M | 560096 | 1 |
| 30 | M | 560190 | 1 |
| 36 | F | 560091 | 2 |
| 39 | M | 560094 | 3 |
| 42 | F | 560099 | 2 |
| 52 | F | 560298 | 3 |
| 53 | M | 560090 | 2 |
| 61 | M | 560092 | 3 |

TABLE III. SAT 1 TABLE CONSIDERING MARITAL STATUS AS A PSA

| GroupID | MaritalStatus | Count |
|---|---|---|
| 1 | Unmarried | 2 |
| 1 | Married | 1 |
| 1 | Unmarried | 2 |
| 2 | Unmarried | 2 |
| 3 | Married | 2 |
| 2 | Married | 1 |
| 3 | Married | 2 |
| 2 | Unmarried | 2 |
| 3 | Unmarried | 1 |

TABLE IV. SAT 2 TABLE CONSIDERING SALARY AS A CSA

| GroupID | Salary |
|---|---|
| 1 | 45K |
| 1 | 48K |
| 1 | 30K |
| 2 | 55K |
| 3 | 57K |
| 2 | 65K |
| 3 | 45K |
| 2 | 45K |
| 3 | 40K |

TABLE V. SAT 3 TABLE CONSIDERING DISEASE AS A CSA

| GroupID | Disease |
|---|---|
| 1 | Flu |
| 1 | Pneumonia |
| 1 | Flu |
| 2 | Bronchitis |
| 3 | Heart Infection |
| 2 | Heart Attack |
| 3 | Gastric Ulcer |
| 2 | Dyspepsia |
| 3 | Heart Infection |

From the generated tables it can be observed that marital status is considered to be HAS and the distribution of all the records was done based on this attribute. In Table III, the first EQ contains two occurrences of attribute value " Unmarried" and one occurrence of "Married". In Table V, the diseases in the EQ1, 'flu' and 'pneumonia' belong to chest infection and the intruder with some background knowledge( age, gender and zip code) can easily get to know the sensitive information. For example, if a person is neighbor of Alice and knows her

QID's , on getting access to the published Tables II, III, IV and V, he concludes that the Alice record belong to EQ1 and that she is suffering from some chest infections. This happened because in EQ1, all diseases are semantically similar.
It can be seen that the PSA is chosen based on number of occurrences. However, when the equivalence classes are created the attributes may be grouped such that they are semantically similar, this leads to semantic attacks and also due to multiple sensitive attributes there are every possibility that there could also be association attacks.

The following are the research gaps observed from the background study:

- Among the existing PPDP algorithms for MSA very few discuss how to select the Primary/Secondary/Tertiary sensitive attributes.
- Most of the algorithms do not deal with the residue records- those records that are skewed and do not fit into any of the equivalence classes.

### B. Main Contribution of the Article

The main contributions of this work are:

- To provide an efficient method to select the sensitive attributes.
- Distributing the records within the EQ groups based on parameter 'e'.
- Applying incremental diversity so as to distribute the records appropriately within EQ with minimal residue records and preventing semantic attacks.
- Comparing the performance of the proposed algorithm (changing the primary sensitive attributes) against various parameters like residue percentage, diversity parameter (e) and computation time.

## IV. Proposed Method and Empirical Results

Initially, the semantic hierarchy tree is constructed for all the selected sensitive attributes. For example, if disease, marital status and relationships are considered as sensitive attributes, the semantic hierarchical tree for all these is shown in Fig. 3,4 and 5 respectively.The semantic hierarchical tree for disease attribute, with Disease labelled as root node is at Level 0, the childrens namely Respiratory Disease and Digestive System diseases are at Level 1 and the attributes under these diseases are at level 3 and so on. Similarly, for attribute Marital Status


Fig. 3. Semantic hierarchical tree for disease attribute (Height=3).

there are 3 levels(0,1 and 2) and for Relationship there are 2 levels. Once the semantic hierarchy trees are constructed, those attributes with trees having more number of levels and


Fig. 4. Semantic hierarchical tree for disease attribute (Height=2).


Fig. 5. Semantic hierarchical tree for disease attribute (Height=1).

with more number of child nodes can be selected as Initial Sensitive Attributes(ISA). This selection is essential to achieve optimal diversity of sensitive attributes in each equivalence classes. For example, if Disease is chosen as a ISA, if the equivalence class consist of sensitive values "Flu", "Heart Infection" and "Jaundice", the class satisfies (3, 2) diversity. Here, the equivalence class contains different values as well as the values are semantically far from each other. If Marital status is chosen as a ISA, then it is difficult to achieve (3,2) diversity, we can achieve only (3,1) diversity by repeating one of the values in each equivalence class. If Relationship is chosen as the ISA then it is possible to achieve only 'l' diversity and achieving (l,e) is not viable. A (3, 1) diversity table is shown in Table VI. Here, Disease sensitive attribute is chosen as the ISA

TABLE VI. QID Table of Table I

| Age | Gender | Zip code | GroupID |
|---|---|---|---|
| 23 | F | 560098 | 1 |
| 25 | M | 560096 | 1 |
| 39 | M | 560094 | 1 |
| 36 | F | 560091 | 2 |
| 30 | M | 560190 | 2 |
| 61 | F | 560092 | 3 |
| 53 | M | 560090 | 3 |
| 52 | F | 560298 | 3 |
| 42 | M | 560099 | 2 |

TABLE VII. SA(Disease) Satisfying (3, 1) Diversity

| Disease | Salary | MaritalStatus | Group ID |
|---|---|---|---|
| Flu | 45K | Unmarried | 1 |
| Pneumonia | 48K | Married | 1 |
| HeartInfection | 57K | Married | 1 |
| Bronchitis | 55K | Unmarried | 2 |
| Flu | 30K | Unmarried | 2 |
| Heart Attack | 65K | Married | 2 |
| GastricUlcer | 45K | Married | 3 |
| Dyspepsia | 45K | Unmarried | 3 |
| HeartInfection | 40K | Married | 3 |

After choosing the ISA, it is necessary to choose secondary sensitive attribute, ternary sensitive attribute and so on. This is necessary because if the Table 6 and 7 are published as they are, it may lead to association attack. For example, consider equivalence class 3, here even though the disease attribute satisfies (3, 1 ) diversity, the other associated attributes are predictable. If the intruder knows that a woman is more than 50 years and she is married, he well be easily be able to get to know that the lady belongs to group 3 and suffering from gastric ulcer. Such an attack is known as association attack [34]. These attacks happen in data set with multiple sensitive attributes.

### A. Choosing Secondary and Tertiary Sensitive Attributes

From previous discussions it is clear that as a primary phase of data anonymization it is necessary to assign certain ranks to sensitive attributes. The attributes can be ranked based on the structure of the semantic tree. Those attribute values for which parents are more can be chosen as ISA and marked as rank 1. The next attributes are those with lesser parents as in case of Marital Status these attributes are termed as Subsequent Sensitive attributes (SSA) with rank 2 . Those sensitive attributes for which there are no many unique values and also are numerical in nature, for such attribute's values within each equivalence classes, they can be replaced with the mean of the values. For example, salary attribute, can be replaced with it mean value in each equivalence class The resulting tables generated based on the categorization of multiple sensitive attributes is shown in Table VIII and IX.

TABLE VIII. QID TABLE OF TABLE 1

| Age | Gender | ZipCode | Group ID |
|---|---|---|---|
| 52 | F | 560298 | 1 |
| 25 | M | 560096 | 1 |
| 39 | M | 560094 | 1 |
| 36 | F | 560091 | 2 |
| 30 | M | 560190 | 2 |
| 61 | F | 560092 | 3 |
| 53 | M | 560090 | 3 |
| 23 | F | 560098 | 3 |
| 42 | M | 560099 | 2 |

TABLE IX. ANONYMIZED TABLE BASED ON RANKS OF THE SENSITIVE ATTRIBUTES

| Disease | Salary | MaritalStatus | GroupID |
|---|---|---|---|
| Disease | Salary | Marital Status | Group ID |
| Dyspepsia | 41K | Unmarried | 1 |
| Pneumonia | | Married | 1 |
| HeartInfection | | Married | 1 |
| Bronchitis | 59K | Unmarried | 2 |
| HeartInfection | | Unmarried | 2 |
| Gastritis | | Unmarried | 2 |
| Gastric Ulcer | 43.3K | Married | 3 |
| HeartAttack | | Married | 3 |
| Flu | | Unmarried | 3 |

When implementing the algorithm, the records are recursively reordered to make sure that in every EQ there is high diversity between the ISA values, average diversity between SSA and so on. That is, there will be incremental diversity achieved over the ranks of the sensitive attributes. The algorithm proposed next, takes the microdata table with identifiers, quasi identifiers and sensitive attributes as input. I, Q and S represents the number of identifiers, quasi identifiers and sensitive attributes respectively. The output of the algorithm is the separate QID table and SA table. The algorithm is

**Algorithm 1** Proposed Algorithm

Input :

1) Microdata M( $i_1, i_2..i_I, q_1, q_2, \ldots q_Q, s_1, s_2, s_3 \ldots .s_S$ ).
2) The diversity parameters 'l' and 'e'.
3) Equivalence group size 'k'.

Output :

1) QIT(q1, q2..qn2 )
2) SA( s1,s2,s3....sn3)

1: Classify the attributes within M into identifiers (i1,i2..in1) quasiidentifiers q1, q2,...qn2 ) and sensitive attributes (s1,s2,s3....sn3)
2: Generate the semantic hierarchy tree for the sensitive attributes T( T1,T2..Tn3).
3: Sort T in ascending order based on the depth of the tree.
4: Select the Sensitive attribute with maximum depth as Initial Sensitive Attribute(ISA), the next as Secondary Sensitive Attribute(SSA) and so on.
5: Initialize the groups EG ( G1,G2....Gm), K=k, QIT= $\Phi$ and SA=$\Phi$
6: Place all the records in the temporary dictionary TD.
7: **while** $T \neq Empty$ **do**
   1) Place ti into EG such that the ISA of ti when placed in EG satisfies 'l','e' defined before.
   2) Increment value of K for that EG.
   3) If not satisfied, place the tuples into Residue Dictionary RD and select next tuple.
   4) If size of EG ¿ K break and place ti in next EG.
8: **end while**
9: **while** $RD \neq Empty$ **do**
   1) reiterate the above steps (9-13) to reduce the residual records.
   2) If the SA is numerical, within each EG, replace all the values by the mean.
   3) Separate the SA's and QID's into separate table. Assign the Group ID's for the groups generated.
10: **end while**

implemented in Python language, the results obtained with varying k, number of records, 'l' and choosing different sensitive attributes. This is discussed in the next section.

## V. EXPERIMENTS AND PERFORMANCE EQUATIONS

The algorithm is implemented in Python language using native python data types tuples, dictionary and lists. The use of external libraries such as NumPy and Pandas is avoided since it increases time complexity of the algorithm. The iteration through the tuples is pretty faster when dictionaries are used. The implemented algorithm is tested on the demographic data set obtained from University of California (UCI) machine learning repository [38]. This microdata contains 30162 records. Occupation, Education, Marital Status, Work Class

TABLE X. NUMBER OF UNIQUE VALUES IN EACH OF MSA'S

| Attribute | Occupation | Education | MaritalStatus | Relationship | Race |
|---|---|---|---|---|---|
| No. of unique values | 14 | 16 | 7 | 7 | 5 |

and Race are chosen as MSA's. Age, gender and Zipcode are chosen as QID's. The number of unique values for each of these is shown in Table X.

The following equations are used to compute various performance parameters. The residue percentage is computed as per equation 1.

$$RP = \frac{TotalNumberofRecordsinRD}{TotalNumberOfRecordsinM} * 100 \quad (1)$$

Where RP- Residue Percentage.
RD- Residue Directory.
M- Original Microdata.

The computation time needed to run the code is obtained using equation 2.

$$Computationtime = (endtime - startime) * 1000 \quad (2)$$

Where end time and star time are initialized at the beginning and end of the program respectively with the function time. time( ) that returns number of seconds elapsed since epoch.

$$DiversityOfEachAttributeWithinAnEQDEA = \frac{Numberofuniquevaluesofattribute}{Totalnumberofvaluesofattributes} \quad (3)$$

The diversity percentage of the entire table is computed using equation 4.

$$DiversityPercentage = \frac{\sum_{EQ=1}^{n} \frac{\sum_{DEA}}{m}}{n} \quad (4)$$

Where n- Total number of EQ's constructed. m- Total number of attributes

## VI. RESULTS AND DISCUSSIONS

The results obtained after performing the experiments is presented in this section. The first three experiments are by varying the primary sensitive attributes and k, observing the residue percentage, computation time and diversity. The next set of experiments discusses the performance of the proposed algorithm with (l, e) diversity algorithm, in terms of residue percentage and computation time. Finally the comparison is done with proposed method, (l, e) diversity [17] and SDSV algorithm [29].

### A. Performance of the Proposed Algorithm

*1) Percentage of residue records based on choosing different primary sensitive attributes:* The main objective is to reduce the residue percentage. Choosing k=3, and records 1000-5000, each line indicates number of residue records left out when a particular attribute is chosen as a ISA. It can be observed in Fig. 6, that if race is chosen as ISA, the percentage of residue records is highest and it is lowest when education is chosen as the ISA. The percentage of residue records is computed as per equation 1.

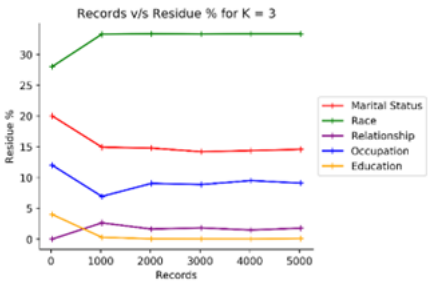


Fig. 6. Percentage of residue records vs parameter K.

*2) Computation time:* The computation time is the time required to generate the final QID and SAT tables. For this, on the chosen number of records, the equivalence classes are to be created choosing the diversity parameter 'l', of l-diversity [12] the levels of sensitive attributes and group size k as defined in [8]. On experimentation it is observed that, when Education is chosen as a ISA it consumes more time than Occupation or Race. This is obvious because the unique values are more for education attribute. The time performance choosing different attributes is shown in Fig. 7. The computation time is as per equation 2.

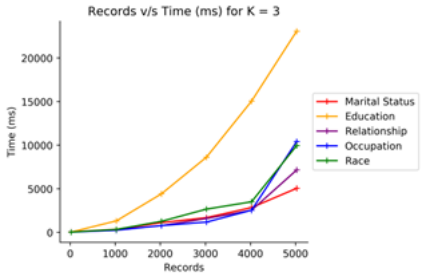


Fig. 7. Time performance for various attributes.

*3) Diversity among the attribute values within the equivalence class:* The diversity is computed as per (l,e) diversity discussed previously. From the experiments it can be observed in Fig. 8, that the attribute with more unique values (Education) achieves better diversity among the attributes within the equivalence classes. With the value of 'k' from 5 to 8, the performance of achieving more diversity can be seen with "education" attribute. The diversity percentage is computed according to equation given in 3 and 4.

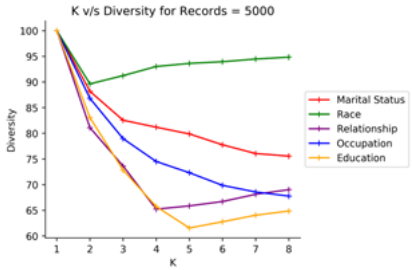


Fig. 8. Diversity of records within each EQ's.

## B. Comparing with (l,e) Diversity Algorithm

The performance of our proposed algorithm is compared with the existing (l,e) diversity algorithm. The (l,e) diversity chooses only one sensitive attribute i.e Education. On observation it can be seen that choosing multiple sensitive attributes and then diversifying records achieves better performance in terms of reducing residue percentage. However, the time taken is more since multiple attributes are considered.

*1) Residual percentage:* The comparison is done for No.of records vs residue records and value of k. It can be observed from Fig. 9 that our proposed algorithm- choosing the attributes based on the ranks and then anonymizing results in reduction of residue records. Since (l,e) diversity uses generalization for anonymization it leads to more number of residue records.

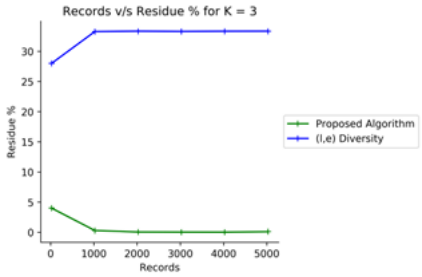


Fig. 9. Reduced residual records in proposed method vs (l, e) diversity.

*2) Computational time:* As shown in Fig. 10, the time taken by the proposed algorithm is slightly higher than (l, e) diversity because the algorithm considers MSA where as (l,e) preserves privacy of SSA.

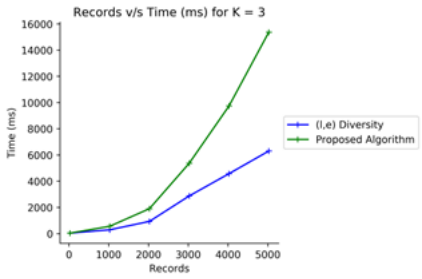


Fig. 10. Computation time in proposed algorithm vs (l, e) diversity.

*3) Diversity percentage:* The diversity percentage achieved in the proposed method with multiple sensitive attributes is much better when compared with (l,e) diversity. This is mainly because the attributes are selected based on their semantics and every EQ has diversified primary sensitive attribute. This is shown in . 11.

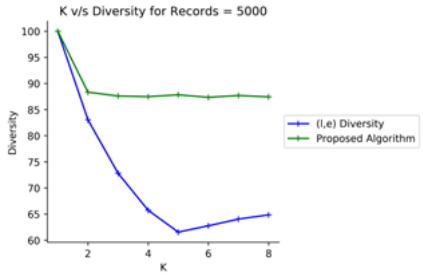


Fig. 11. Diversity percentage in proposed algorithm vs (l,e) diversity.

## C. Comparision of Incremental Diversity, SDSV and (l,e) Diversity

As discussed in related work section, one recent algorithm that discusses the distribution of sensitive attributes is SDSV algorithm. However, the algorithm doesn't consider the semantic similarity between the attributes within an EQ. This leads to semantic diversity attack and weaker diversity among the attributes within an EQ's. It can be seen from Fig. 12, that, (l, e) diversity has highest diversity among the attributes within an EQ, since it considers single sensitive attributes. The diversity percentage for the proposed algorithm is average considering the multiple sensitive attributes and their semantic.

## D. Security Evaluations

As mentioned before the privacy attacks considered in this work are semantic attacks, similarity attacks and association attacks that are predominant in data set with MSA. The proposed algorithm overcomes all these threats since the semantics of the sensitive attributes is addressed. Consider Table II, III, IV and V that were generated by SDSV algorithm. The algorithm did not consider the semantic relationship between the sensitive attributes there were semantically similar attribute values for disease within an equivalence class. Also, the algorithm generates multiple tables and this increases as the number of sensitive attributes increases.

The proposed algorithm overcomes the semantic and similarity attacks. Consider Tables VI and VII that are generated using the proposed algorithm. Every equivalence class has diversity of the sensitive attributes, which becomes difficult for the intruder to cause privacy threats. Even though the intruder knows one of the sensitive attribute and a quasi identifier it is difficult to cause association attack. For example, if the intruder is neighbour of Trudy (from Table I), he knows that he is unmarried and also the Zip Code. The intruder wants to determine other sensitive attributes like disease. On observing the published table, he gets to know that his record belongs to group 2 of Table VII. Here, since there are 2 records that have "Unmarried" as attribute value of Martial Status, he cannot predict Trudy's disease. Form the above experiments

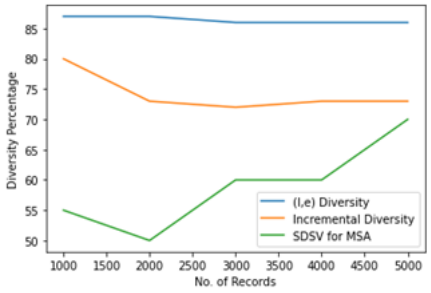


Fig. 12. Diversity percentage of (l, e) diversity, proposed method and SDSV for MSA.

and results it can be concluded that the proposed algorithm is efficient in terms of reducing the residual records, computation time at the same time achieving optimal diversity considering multiple sensitive attributes. Also, as discussed,the proposed algorithm overcomes the privacy threats that exists for MSA.

## VII. Conclusion and Future Work

Concern to data privacy is increased with the increase in the digital technology. The personal data is collected at various places that contain multiple sensitive attributes (MSA). These attributes must be treated well to prevent privacy threats when the data set is published to outside world. Many algorithms have been proposed to preserve privacy of MSA in the literature. In these algorithms one of the attribute is chosen as a primary sensitive attribute and the microdata is anonymized. These algorithms do not discuss how to rank the sensitive attributes. This is the essential step in anonymizing the data. In this paper we discuss an efficient approach to rank the sensitive attributes and then anonymize the data. Experiments along with performance parameters, prove that our algorithm outperforms the existing methods and can be efficiently used to anonymize the data. As a part of future work we would propose an infrastructure framework where in the tables can be published.

## References

[1] Quach, S., Thaichon, P., Martin, K.D. et al., "Digital technologies: tensions in privacy and data", J. of the Acad. Mark. Sci. vol.50, pp. 1299–1323,2022.

[2] Fung, Benjamin CM, et al. "Privacy-preserving data publishing: A survey of recent developments", ACM Computing Surveys (Csur), vol.42, issue 4, pp. 1-53, 2010.

[3] Cox, Lawrence H. "Suppression methodology and statistical disclosure control", Journal of the American Statistical Association, vol.75, issue 370, pp. 377-385,1980.

[4] Sowmyarani C. N. and Dayananda P., "Analytical Study on Privacy Attack Models in Privacy Preserving Data Publishing," pp. 98–116. doi: 10.4018/978-1-5225-1829-7.ch006.

[5] L. Sweeny, "Achieving K-Anonymity Privacy Protection Using Generalization and Suppression", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 571–588, Oct. 2002, doi: 10.1142/S021848850200165X.

[6] K. El Emam and F. K. Dankar, "Protecting privacy using K-anonymity," Journal of the American Medical Informatics Association, vol. 15, no. 5, pp. 627–637, 2008.

[7] V. Ciriani, S. D. C. Vimercati, S. Foresti, and P. Samarati, "k - Anonymity", Privacy-Preserving Data Mining. Advances in Database Systems, vol 34. Springer, Boston, MA, 2008, doi: 10.1007/978-0-387-70992-5.

[8] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, " Worst-case background knowledge for privacy-preserving data publishing" Proceedings of 23rd International Conference on Data Engineering, 126–135, 2007

[9] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," ACM Trans Knowl Discov Data, vol. 1, no. 1, 2007, doi: 10.1145/1217299.1217302.

[10] Ninghui Li, Tiancheng Li, and S. Venkatasubramanian, "Closeness: A New Privacy Measure for Data Publishing," IEEE Trans Knowl Data Eng, vol. 22, no. 7, pp. 943–956, Jul. 2010, doi: 10.1109/TKDE.2009.139

[11] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," VLDB 2006 - Proceedings of the 32nd International Conference on Very Large Data Bases, pp. 139–150, 2006,doi: 10.5555/1182635.1164141

[12] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A new approach for privacy preserving data publishing," IEEE Trans Knowl Data Eng, vol. 24, no. 3, pp. 561–574, 2012, doi: 10.1109/TKDE.2010.236.

[13] D. Li, X. He, L. bin Cao, and H. Chen, "Permutation anonymization", Journal of Intellegent Information Systems,47(3) 427–445, 2016.

[14] He, X., Xiao, Y., Li, Y., Wang, Q., Wang, W., Shi, B: "Permutation Anonymization: Improving Anatomy for Privacy Preservation in Data Publication", Cao, L., Huang, J.Z., Bailey, J., Koh, Y.S., Luo, J. (eds) New Frontiers in Applied Data Mining. PAKDD 2011. Lecture Notes in Computer Science, Springer, Berlin, 7104, 111-123,2012.

[15] M. Bahrami and M. Singhal,"A light-weight permutation based method for data privacy in mobile cloud computing", Proceedings - 2015 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, MobileCloud, pp. 189–196,2015.

[16] K. Oishi, Y. Sei, Y. Tahara, and A. Ohsuga, "Semantic diversity: Privacy considering distance between values of sensitive attribute", Computer Security,94, 1–49,2020.

[17] H. Wang, J. Han, J. Wang, and L. Wang "(l, e)-Diversity – A Privacy Preserving Model to Resist Semantic Similarity Attack", Journal of Computers,9(1),59–64, 2014.

[18] Aggarwal, C. C.,"On k-anonymity and the curse of dimensionality", VLDB, 5, 901-909,2005.

[19] X.-C. Yang, Y.-Z. Wang, B. Wang, and G. Yu,"Privacy Preserving Approaches for Multiple Sensitive Attributes in Data Publishing", Chinese Journal of Computers,31(4),574–587,2009.

[20] A. Anjum, N. Ahmad, S. U. R. Malik, S. Zubair, and B. Shahzad, "An efficient approach for publishing microdata for multiple sensitive attributes",Journal of Super Computers, 74(10), 5127–5155,2018.

[21] F. Liu, Y. Jia, and W. Han, "A new k-anonymity algorithm towards multiple sensitive attributes", In:Proceedings of IEEE 12th International Conference on Computer and Information Technology, CIT 2012,768–772,2012.

[22] T. S. Gal, Z. Chen, and A. Gangopadhyay, "A Privacy Protection Model for Patient Data with Multiple Sensitive Attributes", International Journal of Information Security and Privacy (IJISP), 2(3), 28–44, 2008.

[23] T. Yi and M. Shi, "Privacy Protection Method for Multiple Sensitive Attributes Based on Strong Rule", Mathematical Problems in Engineering, 2015, 1024-123X, 2015.

[24] T. Kanwalet al,"A robust privacy preserving approach for electronic health records using multiple dataset with multiple sensitive attributes", Computer Security, 105,102224, 2021.

[25] R. Wang, Y. Zhu, T.-S. Chen, and C.-C. Chang, "Privacy-Preserving Algorithms for Multiple Sensitive Attributes Satisfying t-Closeness", Journal of Computer Science Technology, 33(6) 1231–1242, 2018.

[26] Widodo, M. Nugraheni, and I. P. Sari, "Simple Distribution of Sensitive Values for Multiple Sensitive Attributes in Privacy Preserving Data Publishing to Achieve Anatomy", Proceedings of 2nd International Conference on Innovative and Creative Information Technology(ICITech 2021),216–220,2021.

[27] LeFevre, K., DeWitt, D. J., Ramakrishnan, R,"Mondrian multidimensional k-anonymity",Proceedings of 22nd International conference on data engineering (ICDE'06), 25-25,2006.

[28] T. M. Truta and B. Vinay, "Privacy Protection: p-Sensitive k-Anonymity Property", Proceedings of 22nd International Conference on Data Engineering Workshops (ICDEW'06), 94–94, 2006.

[29] X. Sun, H. Wang, T. M. Truta, J. Li, and P. Li: (p+,$\alpha$)-sensitive k-anonymity: A new enhanced privacy protection model. In: Proceedings of IEEE 8th International Conference on Computer and Information Technology(CIT 2008) 59–64 (2008).

[30] CN Sowmyarani, Veena Gadad, Dayananda P, "(p+, $\alpha$, t)-Anonymity Technique Against Privacy Attacks", International Journal of Information Security and Privacy (IJISP),15(2),68–86,2021.

[31] J. Liu, J. Luo, and J. Z. Huang," Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity Requirements",Proceedings of 11th International Conference on Data Mining Workshops,666–673,2011.

[32] Y. Wu, X. Ruan, S. Liao, and X. Wang, "P-cover k-anonymity model for protecting multiple sensitive attributes",Proceedings of 5th International Conference on Computer Science and Education( ICCSE 2010), 179–183, 2010).

[33] Y. Ye, Y. Liu, C. Wang, D. Lv, and J. Feng,"Decomposition: Privacy preservation for multiple sensitive attributes",Lecture Notes in Computer Science (including sub-series Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5463,486–490,2009).

[34] D. Das and D. K. Bhattacharyya,"Decomposition+: Improving l-diversity for Multiple Sensitive Attributes", Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST. 85(2),403–412,2012.

[35] S. A. Onashoga, B. A. Bamiro, A. T. Akinwale, and J. A. Oguntuase,"KC-Slice: A dynamic privacy-preserving data publishing technique for multisensitive attributes", Information Security Journal.26(3) 121–135, 2017.

[36] N. V. S. Lakshmipathi Raju, M. N. Seetaramanath, and P. Srinivasa Rao,"An enhanced dynamic KC-slice model for privacy preserving data publishing with multiple sensitive attributes by inducing sensitivity", Journal of King Saud University - Computer and Information Sciences,34(1),1394-1406,2018.

[37] J. Jayapradha, M. Prakash, Y. Alotaibi, O. I. Khalaf, and S. A. Alghamdi, "Heap Bucketization Anonymity—An Efficient Privacy-Preserving Data Publishing Model for Multiple Sensitive Attributes", IEEE Access,10, 28773–28791, 2022.

[38] D. W. Murphy, P. M., and Aha: UCI repository of machine learning databases. https://archive.ics.uci.edu/ml/datasets/adult,1996.