**ABSTRACT**

Myriads of privacy preserving algorithms are present but very few aim to take into consideration the privacy for multiple sensitive attributes. Many pre-existing algorithms aim to segregate the table into sensitive and non-sensitive tables (Anatomy [1]) or completely mask or generalize the information (k-anonymity [30], l, e diversity [16]). This paper combines the best of both worlds and develops an innovate algorithm termed as incremental diversity.

Incremental diversity algorithm is able to diversify the data for multiple sensitive attributes and also produce lesser quantity of residue records as compared to l, e diversity [16]. Incremental diversity mainly chooses a primary sensitive attribute having a certain set of favorable characteristics such the presence of a greater number of unique values as compared to other sensitive attributes, many parents in its semantic hierarchical tree, and also more varying frequency for each sensitive value in the dataset. This paper uses the concept of Anatomy [1] to produce the quasi-identifier and sensitive tables (QIT & ST) along with a masked microdata table using the concept of k-anonymity [30] and l diversity [16]. Either of these sets of tables could be produced as the final table depending on its end-case use scenario and private information sensitivity.

The inference obtained after careful examination of the graphs is that incremental diversity has much better time complexity, generates lesser number of residue records for a given value of k [30], but it comes at the cost of having lesser diversity than l, e diversity. The trade-off between diversity and residue records produced for a given dataset can be adjusted depending on the algorithm’s use-case and the degree of sensitivity that is desired.

**1. INTRODUCTION:**

If the data is published in a highly hidden and inconspicuous way in order to protect privacy, the researchers who need the data cannot extract much information.For example, consider the table showing data of students:

**Table 1.1: Sample student data**

|  |  |  |  |
| --- | --- | --- | --- |
| Month of birth | Hall ticket number | Percentage of marks | College name |
| February | 5243 | 81% | P |
| February | 1673 | 92% | Q |
| February | 5882 | 88% | Q |
| July | 3227 | 61% | R |

Let A be one among the students whose data is displayed. An attacker can easily get to know A’s college name by seeing the above table (Table 1.1). Imagine, the above data is published in the manner subsequently mentioned in order to protect privacy of students.

|  |  |  |  |
| --- | --- | --- | --- |
| Month of birth | Hall ticket number | Percentage of marks | College name |
| January - April | 1000 -8000 | 60 - 100% | P |
| January - April | 1000 - 8000 | 60 - 100% | Q |
| January - April | 1000 - 8000 | 60 - 100% | Q |
| May - August | 1000 - 8000 | 60 - 100% | R |

**Table 1.2: Modified student data**

From the above table (Table 1.2), attacker cannot get to know A’s college name. **S**uppose a researcher or data analyst wants to know how many students from college P have scored more than 80%, then the researcher cannot get to know from the above data. Now, with the help of above table, two tables can be made.

**Table 1.3: Quasi-identifier table (QIT)**

|  |  |  |  |
| --- | --- | --- | --- |
| Month of Birth | Hall ticket number | Percentage of marks | Group ID |
| February | 5243 | 81% | 1 |
| February | 1673 | 92% | 1 |
| February | 5882 | 88% | 1 |
| July | 3227 | 61% | 2 |

**Table 1.4: Sensitive table (ST)**

|  |  |  |
| --- | --- | --- |
| Group ID | College name | Count |
| 1 | P | 1 |
| 1 | Q | 2 |
| 2 | R | 1 |

The researcher can find that one student from college P has scored more than 80% as Group ID 1 corresponds to P and Q and count of P is one (seeing second table), so by observing first table, 81%, 92% and 88% corresponds to either P or Q. As all three scores are above 80%, it can be understood that one student from college P has scored above 80% and college name of P is hidden. This method is called Anatomy [1] (discussed in Background and Previous Works Section).

**2. BACKGROUND AND PREVIOUS WORKS**

The previous researches in the process of publishing the privacy preserved data is mainly focused on single sensitive attribute and which cannot be implemented for the multiple sensitive attributes. The model of k-anonymity [30] was based on partitioning the given microdata table in a manner such that each equivalence class of the microdata table should contain at least k number of records which cannot be distinguishable from the other set of records. The drawback of this model is inefficient resistance to homogenous attacks.

[11] describes the model of preserving the privacy of the data using p-sensitive model for multiple sensitive attributes. This model assigns the Group ID to equivalence classes along with masking and generalizing the quasi-identifier group. The drawback is that it does not take into account the frequency of multiple sensitive attributes.

[16] *(l-diversity)* describes the method of grouping the data. There should be atleast *l* different sensitive values. If the attacker wants to identify the individual with particular sensitive values, then he must have at least *(l-1)* sensitive values with high probability of occurring. The drawback of this model is that it applies only to single sensitive attribute.

[13] describes the method of anonymizing the data based on multiple sensitive attributes. l-m-d anonymity deals with method of anonymizing the microdata table by creating equivalence classes along with generalizing and masking. The drawback of l-m-d anonymity is that it takes into account the process of generalizing the quasi-identifier.

[27] deals with the problem of generalization by dividing the microdata table into the two tables namely quasi-identifier table and the sensitive attribute table. In the quasi-identifier table, the quasi-identifiers pertaining to an individual is displayed along with their Group ID. The sensitive attribute table consists of the Group ID, sensitive attribute and the count of the sensitive value. The process of anatomy was strong and effective but there was a need for diversifying data.

[6] In the process of converting the data into more diversified form, it is important to identify the primary and secondary sensitive attribute. The sensitive attribute containing higher and more unique sensitive values is considered as primary sensitive attribute.

[14] describes the method of calculating the diversity index for the anonymized table. The degree of diversity is an indication of how diversified the table is.

**3. BASIC DEFINITIONS**

**Microdata table:** A raw data obtained from any survey or census (for example: details of all students in a class)

**Quasi Identifier attribute (QI):** An attribute which in combination with other attributes can identify an individual uniquely. (For example: gender, age etc.)

**Sensitive attribute (SA):** An attribute the individual wants to maintain secret or private (for example: his marital details, bank account number etc.)

**Equivalence class:** A table of records consisting of above-mentioned attributes i.e., quasi-identifier attribute and sensitive attribute which is obtained from microdata table. It is basically a subclass obtained from microdata table.

**Semantic hierarchical tree:** A representation in the form of a tree which describes the relationship between various values or attributes (similar to family tree of a family).

**4. PROPOSED METHODOLOGY**

I. Standardization of Data: The microdata table is imported in a standardized manner and stored with the implementation of a nested dictionary. The group id or more specifically, the equivalence class number is assigned to each record with the formula given below.

II. Diversification: Five different algorithms are compared in the program which lead to a variety of intricate results which will be covered in the Results section of the paper. The general methodology employed all five diversification algorithms:

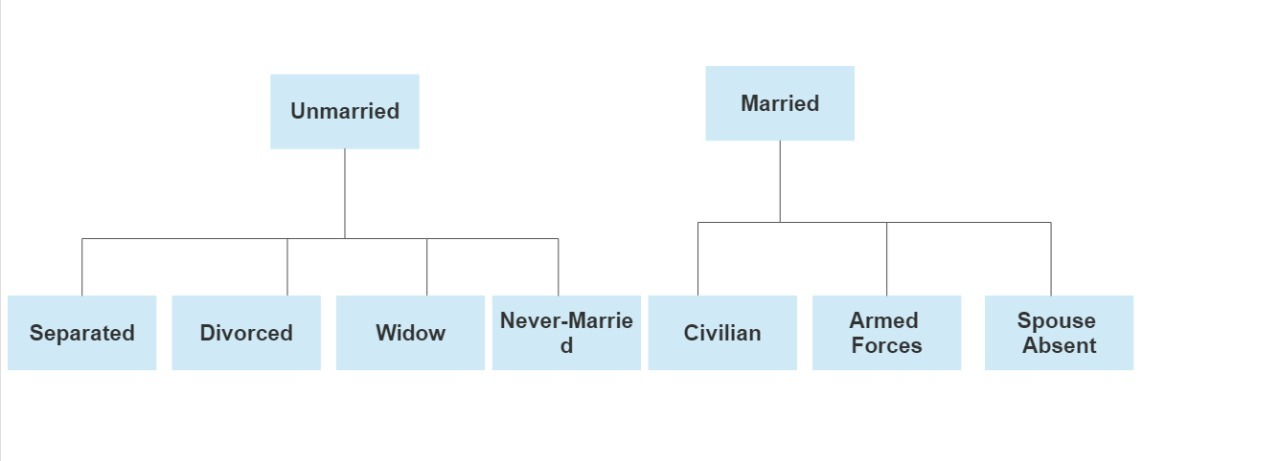
1. The creation of a separate dictionary (called the temporary dictionary) to store the records that do not fit in the selected equivalence class of the microdata table on the basis of the conditions associated with different diversification algorithms.

2. Picking and placing records from the temporary dictionary into the new microdata dictionary based on a particular set of parameters along with the extra condition that the size of the existing equivalence class should be strictly lesser than k (k-anonymity [30]).

3. The diversity of Secondary, Tertiary and Quaternary Sensitive Attributes is increased by swapping their sensitive values which are repeating in a particular equivalence class with non-repeating values from the temporary dictionary.

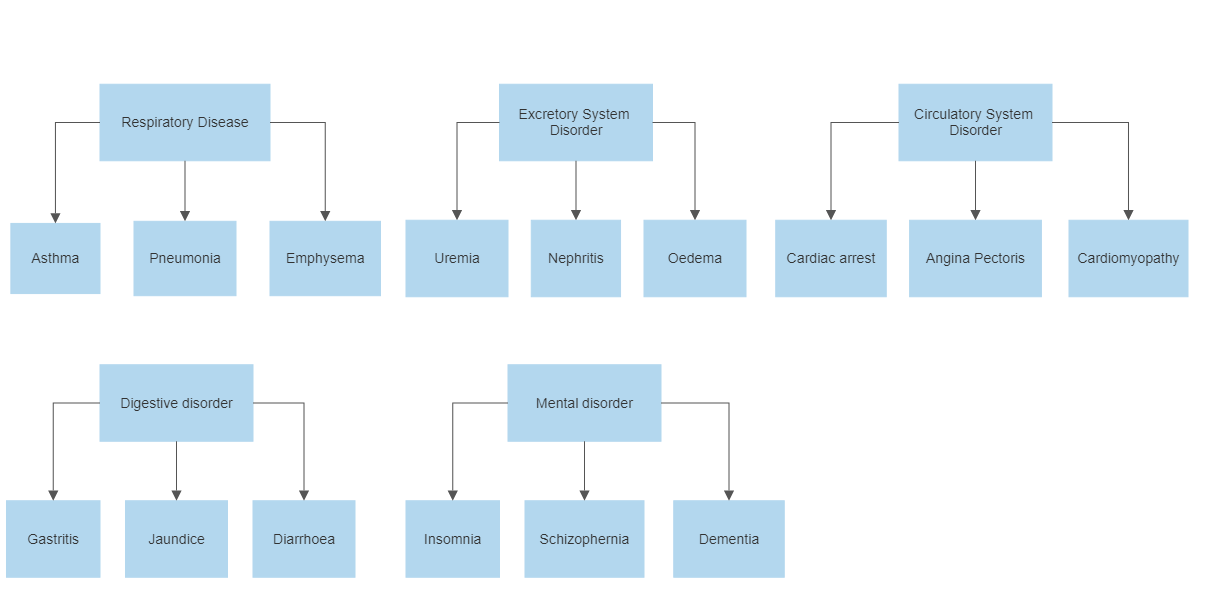
The five diversification algorithms on the basis of which records are removed from the microdata table, added to the temporary dictionary, and added back to the modified microdata table are:

1. Based on the unique appearance of a primary sensitive attribute with few parents in its semantic hierarchical tree (Eg: Marital Status): In a particular equivalence class, a value of the chosen primary sensitive attribute appears only once, i.e., the sensitive value is unique for a specific equivalence class.
2. Based on the unique appearance of a parent in the semantic hierarchical tree consisting of a fewer number of parents for the primary sensitive attribute: It is similar to the former algorithm with the key difference being that instead of the sensitive value, the parent of the chosen primary sensitive attribute appears only once in the group, i.e., it is unique for a specific equivalence class. Example: If a record in an equivalence class has the value of “Widow” for the primary sensitive attribute “Marital Status”, the parent of “Widow” being “Unmarried”, no other record containing Marital Status value as Separated, Divorced or Never-Married is allowed to be added into the equivalence class.
3. Based on the appearance of a parent in the semantic hierarchical tree consisting of a fewer number of parents for the primary sensitive attribute at most twice: The parent of the chosen primary sensitive attribute is allowed to appear twice in a particular equivalence class.



**Figure 4.1: Semantic Hierarchical Tree for Marital Status**

1. Based on the unique appearance of a primary sensitive attribute with more varying frequency of each value in the dataset (Eg: Relationship): It is an offshoot of the first algorithm where only the primary sensitive attribute is changed from one having fewer parents in its semantic hierarchical tree (Marital Status) to the attribute having more variations in its values (Relationship).
2. Based on the unique appearance of a parent in the semantic hierarchical tree consisting of high number of parents for the primary sensitive attribute: A primary sensitive attribute with a higher no. of parents in its semantic hierarchical tree is chosen (Disease) and the parent uniquely appears in an equivalence class. For example, if in an equivalence class there is the presence of a disease for a particular record, let’s say, Cardiomyopathy, belonging to the Circulatory System Disorder Category (Parent), then no other record with the disease belonging to the same parent (Circulatory System Disorder) will be allowed to be placed in that equivalence class.



**Figure 4.2: Semantic Hierarchical Tree for Disease**

III. Segregation of the modified microdata table into Quasi-Identifier Table (QIT) and Sensitive Table (ST).

IV. Masking and generalisation of the attributes with ease & efficiency.

Generalising Age:

This a type of cyclic masking with step size of 3.

Generalising Gender: “M/F” is used.

Masking Zip Code: Only the last three digits are hidden with (\*). (Eg: 560\*\*\*)

Masking Employment, Race or Salary: Asterisk symbol (\*) is used.

V. The various performance parameters to evaluate and analyse the algorithm for varying values of the inputs (no. of records and k) is calculated with the following formulas:

Residue Percentage =

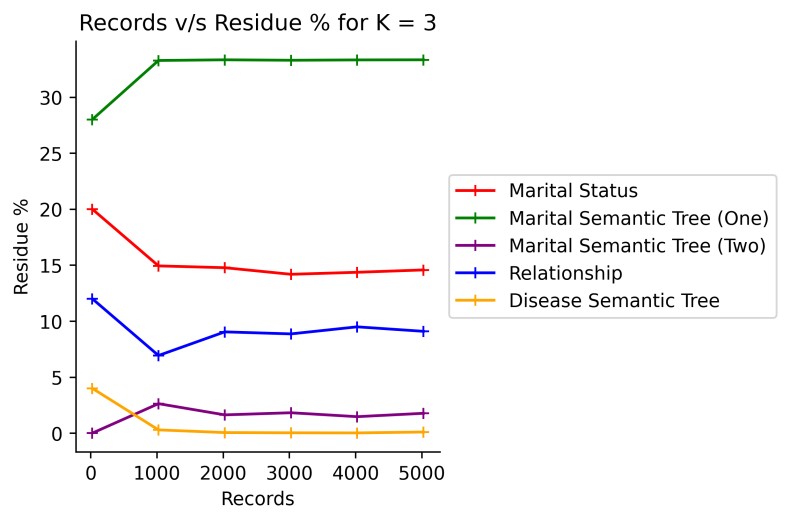
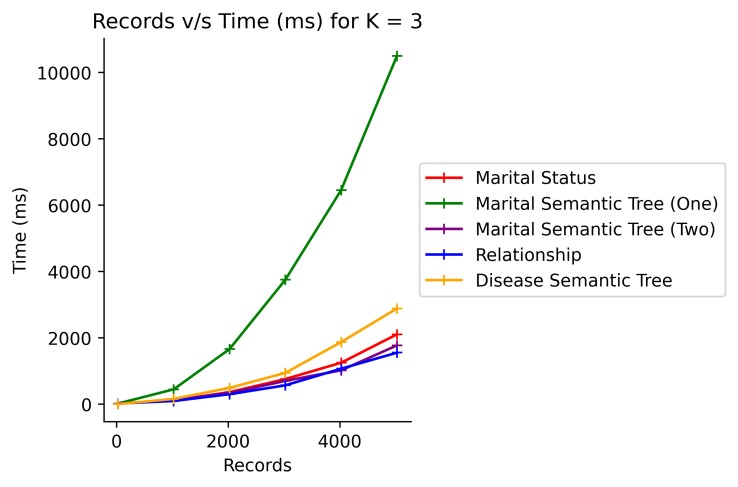
Diversity Percentage =

The diversity for each NMA is calculated as follows:

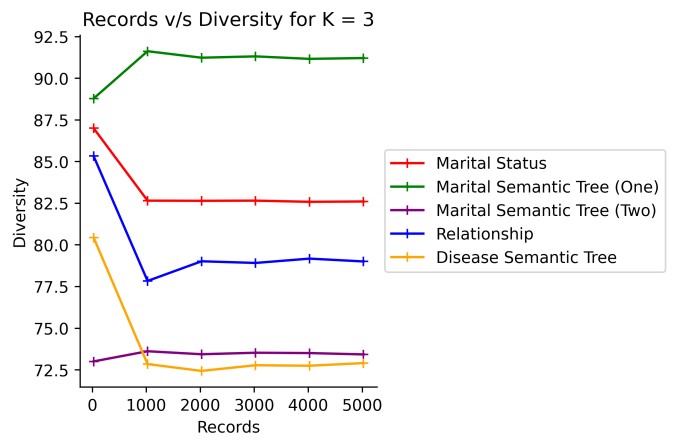
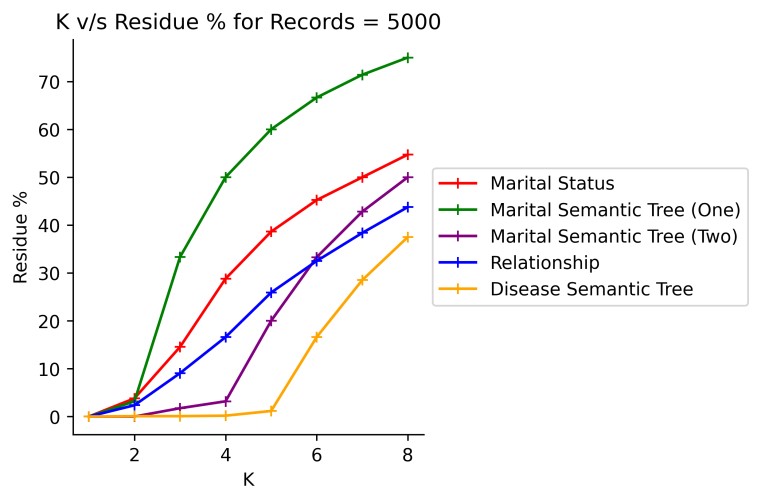
Diversity for each NMA =

**5. EXPERIMENTS**

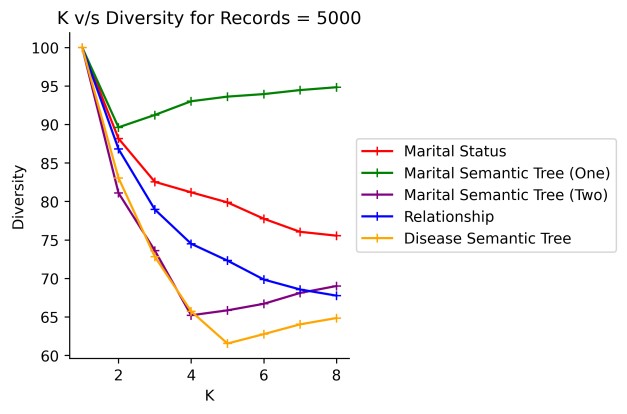
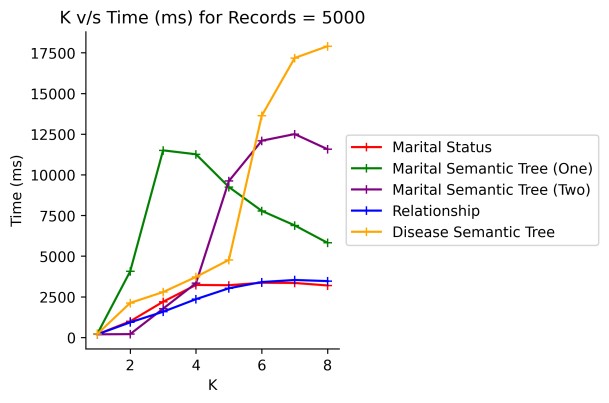
Numerous graphs for various parameters and constant conditions are plotted

**Fig 5.1 Fig 5.2**

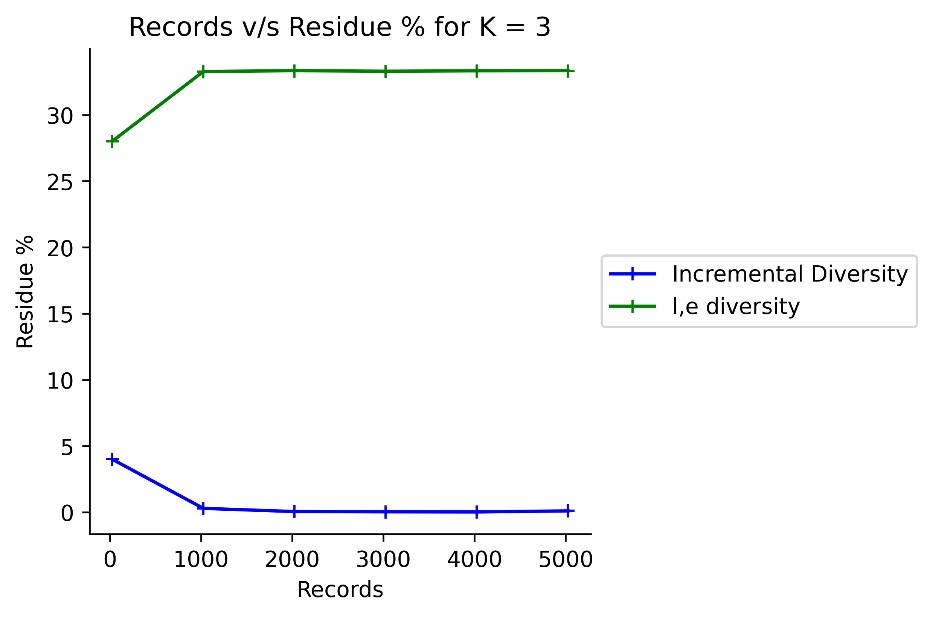
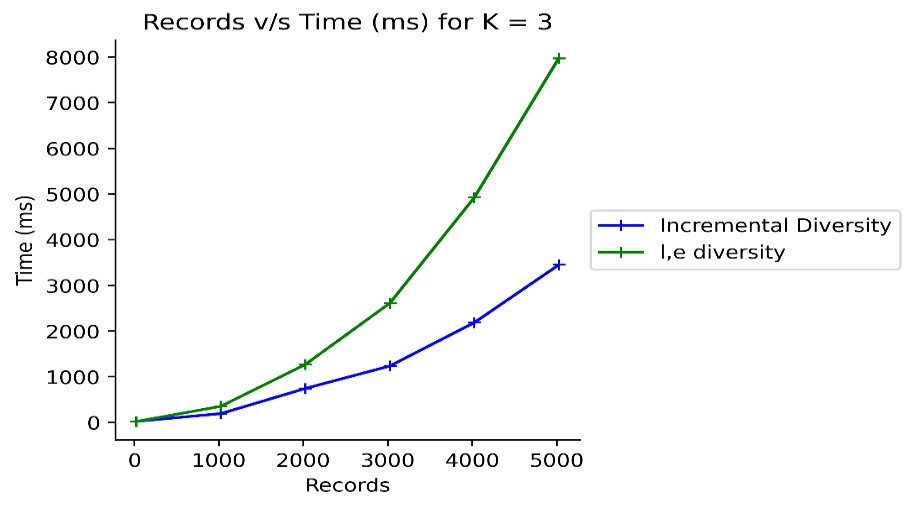
 

**Fig 5.3 Fig 5.4**

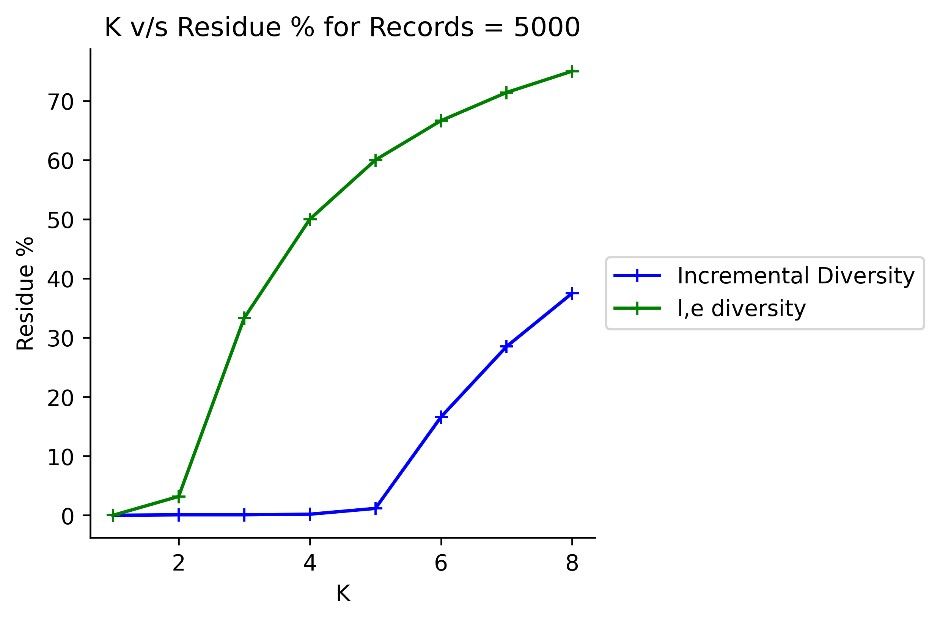
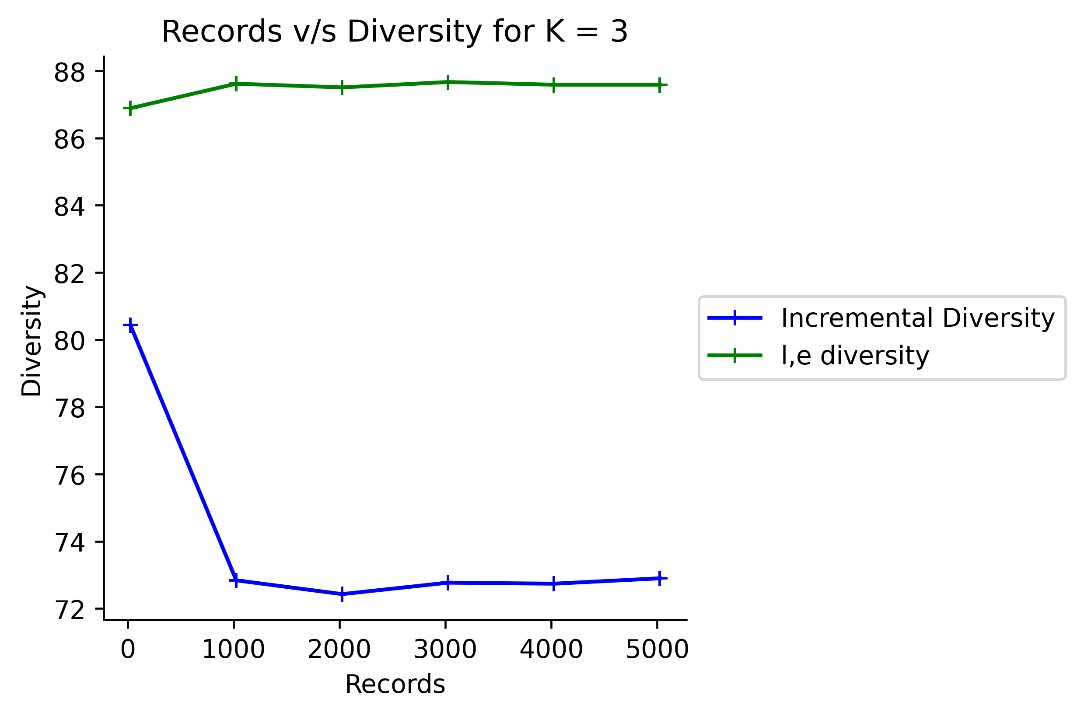


**Fig 5.5 Fig 5.6**

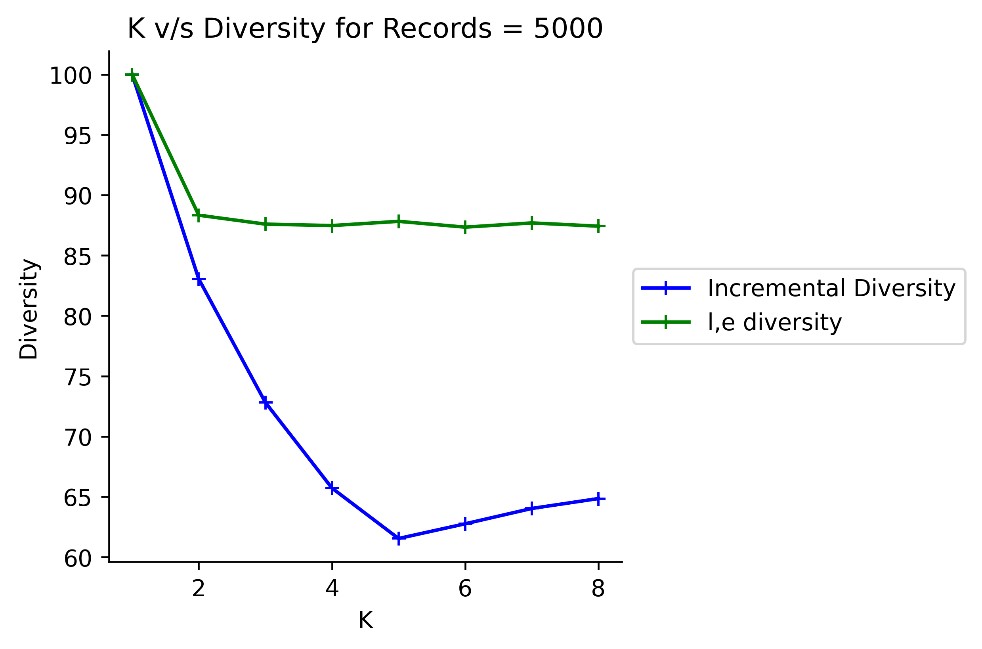
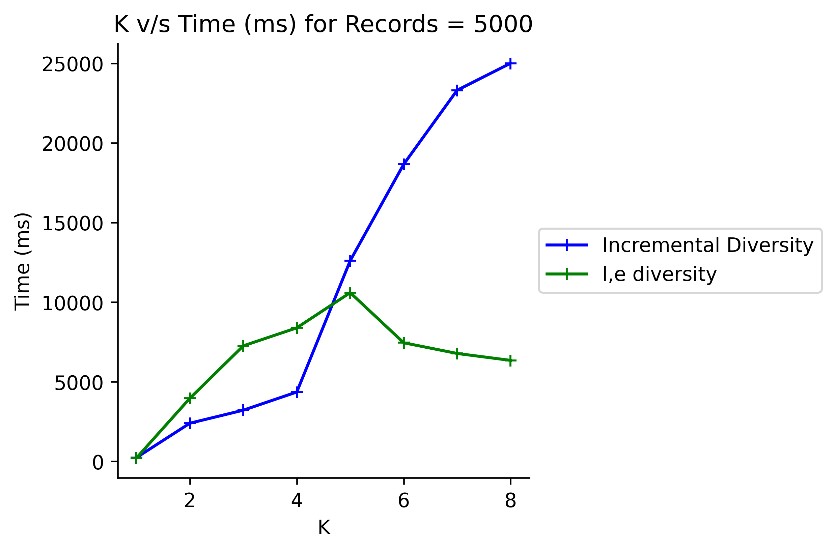
Comparison of l, e Diversity and Incremental Diversity: -



**Fig 5.7 Fig 5.8**



**Fig 5.9 Fig 5.10**



**Fig 5.11 Fig 5.12**

**6. RESULTS**

**1. No. of Records v/s Residue %:** From Fig 5.1, the best possible case for arrangement of records will be based on unique Diseases (algorithm 5) in each equivalence group as the residue % for it is very less for larger records since number of unique Diseases are more than the number of unique Marital Parents.

**2. No. of Records v/s Time (ms):** In Fig 5.2, since the uniqueness of the Marital Parent is very low, large number of residues will be obtained in the residue list. Therefore, more time is spent iterating through the residue list multiple times to replace records.

**3. No. of Records v/s Diversity:** In Fig 5.3, for algorithm 2, the diversity increases as the number of records increases because most of the records will be sent to residue list hence less repetition in the modified microdata table and more diversity.

**4. K v/s Residue %:** In Fig 5.4, there is a steady increase in residue % for all algorithms because when the K value increases there will be more repetition which leads to higher residue %.

**5. K v/s Time (ms):** From Fig 5.5, the time taken for execution increases as the value of K increases, since the program iterates through larger residuals and then swaps the datasets leading to increase in the time taken for execution.

**6. K v/s Diversity:** In Fig 5.6, the diversity is very high for smaller values of K because the no. of unique values is more relative to K values, so less repetition.

**Quasi Identifier Table (QIT) & Sensitive Table (ST)**

**Table 4.13: Quasi – identifier table**

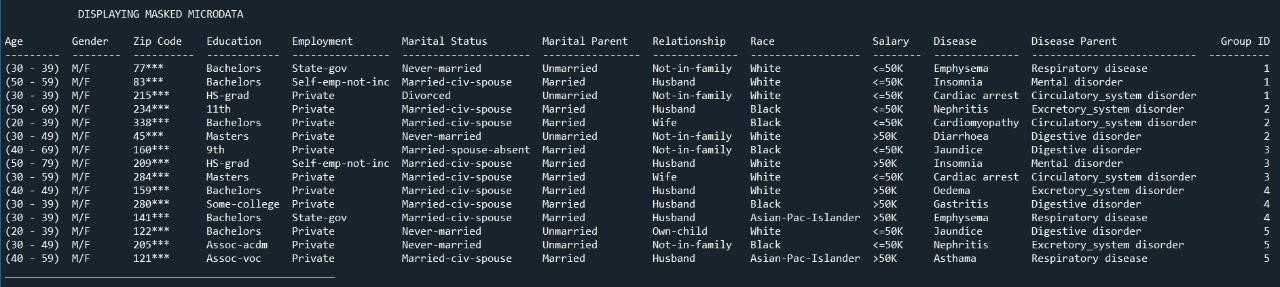


**Table 4.14: Sensitive table (ST) Table 4.15: Count table**



**Masked Microdata Table**

**Table 4.16: Masked microdata table**



**7. Discussion**

**Comparison between l, e diversity algorithm and incremental diversity algorithm**

l, e diversity algorithm employs the use of a primary sensitive attribute with lesser no. of parents in its semantic hierarchical tree (Eg: Marital Status) having one common parent for the sensitive value in the equivalence class.

Incremental diversity algorithm makes use of a primary sensitive attribute with more no. of parents in its semantic hierarchical tree (Eg: Disease) allowing only non-repeating parents to be present in each equivalence class while also performing the incremental diversification for secondary, tertiary and quaternary sensitive attributes. It outperforms l, e diversity in terms of faster time performance and overall decrease in residue records percentage. From the Fig 5.7 and Fig 5.8 we can infer that incremental diversity produces lesser residue records and runs faster in spite of performing diversification for multiple sensitive attributes. Fig 5.9, producing lesser residue records is a double-edged sword which leads to lesser diversity in the records present in the equivalence classes of the table which is a drawback of the incremental diversity algorithm as compared to l, e diversity algorithm.

**8. Conclusion**

Choosing Primary Sensitive Attribute**:** The choice of the right primary sensitive attribute can make or break the diversification and privacy of a table. There will be a trade-off between residue records produced and diversity of the final private table. For example, Disease as the primary sensitive attribute (algorithm 5) produces lesser no. of residue records but at the same time diversity of the records is hampered. On the other hand, Marital Status as the primary sensitive attribute (algorithm 1) produces highly diverse table but it comes at the cost of more no. of residues produced and the loss of precious data. It depends on the use case as to whether diversity needs to be sacrificed in order to produce lesser residue records and prevent huge data loss or, the diversity of the data is of utmost importance and it is okay to overlook the production of enormous residue records.

**REFERENCES**

1. Boyu Li, Yanheng Liu, Minghai Wang, Geng Sun and Bin Li. Local anatomy for personalized privacy protection, 2021.
2. Widodo, Murien Nugraheni and Irma Permata Sari. Simple Distribution of sensitive values for multiple sensitive attributes in privacy preserving data publishing to achieve anatomy, 2021.
3. Khan Razaullah, Tao Xiaofeng, Anjum Adeel, Haider Sajjad and Saif ur Rehman Malik. Privacy Preserving for Multiple Sensitive Attributes against Fingerprint Correlation Attack Satisfying c-Diversity, 2020.
4. Yuelei Xiao and Haiqi Li. Privacy Preserving Data Publishing for Multiple Sensitive Attributes Based on Security Level, 2020.
5. Jassma N Vanasiwala and Nirali R Nanavati. Privacy Preserving Data Publishing of multiple Sensitive Attributes by using Various Anonymization Techniques, 2020.
6. W Widodo and A Wahyudin, A preliminary phase on anatomyzing multiple sensitive attribute by determining main sensitive attribute, 2020.
7. Widodo Widodo, Eko K. Budiardju, Wahyu Catur Wibowo and Harry Achson. An approach for distributing sensitive values in k – anonymity, 2019.
8. Rong Wang, Yan Zhu, Tung-Shou Chen and Chin-Chen Chang. Privacy-Preserving Algorithms for Multiple Sensitive Attributes Satisfying t-Closeness, 2018.
9. Ozgu Can. Personalised Anonymity for micro data release, 2018.
10. Edwin M. Knorr, Giulio Valentino Dalla Riva and Orlin Vakarelov. Anatomy of a new data science course in privacy, ethics and security, 2018.
11. Widodo and Wahyu Catur Wibowo. A distributional model of sensitive values on p – sensitive in multiple sensitive attributes, 2018.
12. Wantong Zheng, Zhongyue Wang, Tongtong Lv and Yong Ma. K – Anonymity Algorithm based on improved clustering, 2018.
13. Junjie Jia and Luting Chen. (l,m,d) – Anonymity : A Resisting similarity attack model for multiple sensitive attributes, 2017.
14. Haiyuan Wang, Jianmin Han and Jivi Wang. (l,e) – diversity A privacy preserving model to resist semantic similarity attack, 2014.
15. Hua Jin, Shang – Cheng Liu and Shi-Guang Ju. Privacy preserving for multiple sensitive attributes based on l – coverage, 2014.
16. Michael Kern. Anonymity: A Formalization of privacy – l – diversity, 2013.
17. Sunyong Yoo, Moonshik Shin and Doheon Lee. An Approach to Reducing Information Loss and Achieving Diversity of Sensitive Attributes in k-anonymity Methods, 2012.
18. Xiang Min Ren, Jing Yang, Jian Pei Zhang and Zong Fu Jia. Uncertain Data Privacy Protection Based on K-Anonymity via Anatomy, 2012.
19. Vijayarani Mohan. K-Anonymity Techniques – A Review, 2010.
20. Khaled El Emam and Fida Kamal Dankar. Protecting Privacy Using k-Anonymity, 2008.
21. Alina Campan and T.M. Truta. Data and Structural k-Anonymity in Social Networks, 2008.
22. Ashwin Machanavajjhala, Daniel Kifer and Johannes Gehrke. L -diversity, 2007.
23. V. Ciriani, S. De Capitani di Vimercati, S. Foresti and P. Samarati. k-Anonymity, 2007.
24. Xioakui Xiao and Yufei Tao. Anatomy: Privacy and Correlation Preserving Publication, 2006.
25. T.M. Truta and Bindu Vinay. Privacy Protection: p-Sensitive k-Anonymity Property, 2006.
26. M. Nergiz and Chris Clifton. Thoughts on k-Anonymization, 2006.
27. Xioakui Xiao and Yufei Tao. Anatomy: Simple and Effective Privacy Preservation, 2006.
28. Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer and Muthuramakrishnan Venkitasubramaniam. ℓ-Diversity: Privacy Beyond k-Anonymity, 2005.
29. Wei Jiang and Chris Clifton. Privacy-Preserving Distributed k-Anonymity, 2005.
30. L. Sweeney. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, 2002.
31. Latanya Sweeney. k- anonymity: A model for protecting privacy, 2001.