RV Educational Institutions ®
RV College of Engineering ®

Autonomous Institution Affiliated to Visvesvaraya Technological University, Belagavi

Approved by AICTE, New Delhi, Accredited By NAAC, Bengaluru And NBA, New Delhi

Go, change the world

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

A NOVEL ALGORITHM FOR PRIVACY PRESERVING DATA PUBLISHING MULTIPLE SENSITIVE ATTRIBUTES

INTERNSHIP REPORT

Submitted by,

1. TANMAY S LAL **1RV21CS176**

2. S MOHAMMED ASHIQ **1RV21CS132**

Under the guidance of

Dr. Sowmyarani C N
Associate Professor
Dept. of CSE
RV College of Engineering

Veena Gadad
Assistant Professor
Dept. of CSE
RV College of Engineering

RESEARCH PROJECT: "Data Anonymization Techniques for Mitigation of Privacy Attacks in Privacy Preserving Data Publishing".

In partial fulfilment for the award of degree
Bachelor of Engineering in
Computer Science and Engineering 2022-2023

RV COLLEGE OF ENGINEERING®, BENGALURU-59

(Autonomous Institution Affiliated to VTU, Belagavi)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the Internship work titled **'A NOVEL ALGORITHM FOR PRIVACY PRESERVING DATA PUBLISHING MULTIPLE SENSITIVE ATTRIBUTES'** is carried out by **TANMAY S LAL (1RV21CS176) and S MOHAMMED ASHIQ (1RV21CS132)** who are bonafide students of RV College of Engineering, Bengaluru, in partial fulfilment for the award of degree of **Bachelor of Engineering in Computer Science and Engineering** of the Visvesvaraya Technological University, Belagavi during the year 2022-2023 for the Summer Internship-I (21CSI310). It is certified that all corrections/suggestions indicated for the Internship have been incorporated in the Internship project report. The Internship report has been approved as it satisfies the academic requirements in respect of internship work prescribed by the institution for the Research Project: Data Anonymization Techniques for Mitigation of Privacy Attacks in Privacy Preserving Data Publishing.

Signature of Guide 1
Dr. Sowmyarani C N

Signature of Head of the Department
Dr. Ramakanth Kumar P

Signature of Guide 2
Veena Gadad

RV COLLEGE OF ENGINEERING®, BENGALURU-59

(Autonomous Institution Affiliated to VTU, Belagavi)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, TANMAY S LAL and **S MOHAMMED ASHIQ** students of second semester B.E., department of CSE, RV College of Engineering, Bengaluru, hereby declare that the Internship Work titled **'A NOVEL ALGORITHM FOR PRIVACY PRESERVING DATA PUBLISHING MULTIPLE SENSITIVE ATTRIBUTES'** has been carried out by us and submitted in partial fulfilment for the award of degree of **Bachelor of Engineering** in **Computer Science and Engineering** during the year 2022-23.

Further we declare that the content of the report has not been submitted previously by anybody for the award of any degree or diploma to any other university.

We also declare that any Intellectual Property Rights generated out of this project carried out at RVCE will be the property of RV College of Engineering, Bengaluru and we will be one of the authors of the same.

Place: Bengaluru

Date:

Name **Signature**

1. TANMAY S LAL (1RV21CS176)

2. S MOHAMMED ASHIQ (1RV21CS132)

ACKNOWLEDGEMENT

We are indebted to our guide**, Dr. Sowmyarani C N (Assosciate Professor) and Veena Gadad (Assistant Professor) , Dept of CSE** for his/her wholehearted support, suggestions and invaluable advice throughout our project work and also helped in the preparation of this report.

Our sincere thanks to **Dr. Ramakanth Kumar P.**, Professor and Head, Department of Computer Science and Engineering, RVCE for his support and encouragement.

We express sincere gratitude to our beloved Principal, **Dr. K. N. Subramanya** for his appreciation towards this project work.

We thank **VGST** for providing us an opportunity to work on this project and enhance our skills.

We thank all the **teaching staff and technical staff** of Computer Science and Engineering department, RVCE for their help.

Lastly, we take this opportunity to thank our **family** members and **friends** who provided all the backup support throughout the project work.

TABLE OF CONTENTS:

ABSTRACT

In this work, we present an Incremental Diversity algorithm that aims to protect the privacy of data while minimizing the loss of data. The algorithm uses the methods of masking and generalization to reduce the number of residue records and prevent significant data loss. The accuracy of the algorithm is evaluated by running aggregate SQL queries on the Google Cloud Platform (GCP) using a dataset of 5000 records. The dataset includes both quasi-identifiers and sensitive attributes, which are vulnerable to threats and attacks by intruders seeking specific sensitive information. The results of the algorithm are compared to the original microdata table, and the effectiveness of the algorithm is demonstrated. The SQL queries and results are explained in detail in the results section . The proposed algorithm can be used to protect the privacy of sensitive data while still allowing for useful data analysis.

INTRODUCTION:

The work presents an Incremental Diversity algorithm that aims to protect the privacy of data by reducing the number of residue records and preventing significant data loss. The accuracy of the algorithm is evaluated by running aggregate SQL queries[1] on the Google Cloud Platform (GCP) using a dataset of 5000 records. The dataset includes both quasi-identifiers (e.g., age, gender) and sensitive attributes (e.g., education, employment, disease) which are vulnerable to threats and attacks by intruders seeking specific sensitive information. The Incremental Diversity algorithm uses the methods of masking and generalization to maintain data privacy. The effectiveness of the algorithm is tested by running SQL queries on GCP, and the results are compared to the original microdata table[2]. The queries and results are explained in detail in the results section.

METHODOLOGY:

The Incremental diversity algorithm aims at protecting the privacy of the data by producing lesser residue records and prevent huge data loss. The accuracy of the algorithm is inspected by running the aggregate SQL queries on the Google Cloud Platform (GCP). The algorithm has been implemented for the dataset containing 5000 records. The sample of the dataset for the first 15 records are shown in the fig.3.1 below.

	Age	Gender	Zip Code	Education	Employment	Marital Status	Marital Parent	Relationship	Race	Salary	Disease	Disease Parent	Group ID
2	39	Male	77516	Bachelors	State-gov	Never-married	Unmarried	Not-in-family	White	<=50K	Emphysema	Respiratory disease	1
3	50	Male	83311	Bachelors	Self-emp-not-inc	Married-civ-spouse	Married	Husband	White	<=50K	Insomnia	Mental disorder	1
4	38	Male	215646	HS-grad	Private	Divorced	Unmarried	Not-in-family	White	<=50K	Cardiac arrest	Circulatory_system disorder	1
5	53	Male	234721	11th	Private	Married-civ-spouse	Married	Husband	Black	<=50K	Nephritis	Excretory_system disorder	2
6	28	Female	338409	Bachelors	Private	Married-civ-spouse	Married	Wife	Black	<=50K	Cardiomyopathy	Circulatory_system disorder	2
7	37	Female	284582	Masters	Private	Married-civ-spouse	Married	Wife	White	<=50K	Cardiac arrest	Circulatory_system disorder	2
8	49	Female	160187	9th	Private	Married-spouse-absent	Married	Not-in-family	Black	<=50K	Jaundice	Digestive disorder	3
9	52	Male	209642	HS-grad	Self-emp-not-inc	Married-civ-spouse	Married	Husband	White	>50K	Insomnia	Mental disorder	3
10	31	Female	45781	Masters	Private	Never-married	Unmarried	Not-in-family	White	>50K	Diarrhoea	Digestive disorder	3
11	42	Male	159449	Bachelors	Private	Married-civ-spouse	Married	Husband	White	>50K	Oedema	Excretory_system disorder	4
12	37	Male	280464	Some-college	Private	Married-civ-spouse	Married	Husband	Black	>50K	Gastritis	Digestive disorder	4
13	30	Male	141297	Bachelors	State-gov	Married-civ-spouse	Married	Husband	Asian-Pac-Islander	>50K	Emphysema	Respiratory disease	4
14	23	Female	122272	Bachelors	Private	Never-married	Unmarried	Own-child	White	<=50K	Jaundice	Digestive disorder	5
15	32	Male	205019	Assoc-acdm	Private	Never-married	Unmarried	Not-in-family	Black	<=50K	Nephritis	Excretory_system disorder	5

Fig.3.1: Original Dataset containing Quasi-identifiers and Sensitive attributes

The Original Dataset containing Quasi-Identifiers (e.g., Age, Gender etc.) and Sensitive attributes (e.g., Education, Employment, Disease etc.) shown above is prone to threats and attacks by the intruder who is searching for the particular sensitive information. In order to maintain the privacy of the data, we have implemented the efficient algorithm called Incremental Diversity. Incremental Diversity algorithm includes the method of masking and generalization. The results of the Original Microdata Table after implementing Incremental Diversity (for the first 15 records) is shown in the fig.3.2. below.

	Age	Gender	Zip Code	Education	Employment	Marital Status	Marital Parent	Relationship	Race	Salary	Disease	Disease Parent	Group ID
2	(30 - 39)	M/F	77***	Bachelors	State-gov	Never-married	Unmarried	Not-in-family	White	<=50K	Emphysema	Respiratory disease	1
3	(50 - 59)	M/F	83***	Bachelors	Self-emp-not-inc	Married-civ-spouse	Married	Husband	White	<=50K	Insomnia	Mental disorder	1
4	(30 - 39)	M/F	215***	HS-grad	Private	Divorced	Unmarried	Not-in-family	White	<=50K	Cardiac arrest	Circulatory_system disorder	1
5	(50 - 69)	M/F	234***	11th	Private	Married-civ-spouse	Married	Husband	Black	<=50K	Nephritis	Excretory_system disorder	2
6	(20 - 39)	M/F	338***	Bachelors	Private	Married-civ-spouse	Married	Wife	Black	<=50K	Cardiomyopathy	Circulatory_system disorder	2
7	(30 - 49)	M/F	45***	Masters	Private	Never-married	Unmarried	Not-in-family	White	>50K	Diarrhoea	Digestive disorder	2
8	(40 - 69)	M/F	160***	9th	Private	Married-spouse-absent	Married	Not-in-family	Black	<=50K	Jaundice	Digestive disorder	3
9	(50 - 79)	M/F	209***	HS-grad	Self-emp-not-inc	Married-civ-spouse	Married	Husband	White	>50K	Insomnia	Mental disorder	3
10	(30 - 59)	M/F	284***	Masters	Private	Married-civ-spouse	Married	Wife	White	<=50K	Cardiac arrest	Circulatory_system disorder	3
11	(40 - 49)	M/F	159***	Bachelors	Private	Married-civ-spouse	Married	Husband	White	>50K	Oedema	Excretory_system disorder	4
12	(30 - 39)	M/F	280***	Some-college	Private	Married-civ-spouse	Married	Husband	Black	>50K	Gastritis	Digestive disorder	4
13	(30 - 39)	M/F	141***	Bachelors	State-gov	Married-civ-spouse	Married	Husband	Asian-Pac-Islander	>50K	Emphysema	Respiratory disease	4
14	(20 - 39)	M/F	122***	Bachelors	Private	Never-married	Unmarried	Own-child	White	<=50K	Jaundice	Digestive disorder	5
15	(30 - 49)	M/F	205***	Assoc-acdm	Private	Never-married	Unmarried	Not-in-family	Black	<=50K	Nephritis	Excretory_system disorder	5

Fig.3.2: Masked Microdata Table after implementing Incremental Diversity algorithm

In order to test the accuracy of the results obtained from the Incremental Diversity algorithm, we run the SQL queries on Google Cloud Platform window (GCP)[2]. The following sections discusses about various queries, results and comparisons.

SQL Queries are the excellent tool for storing, analyzing and retrieving the data in the specific and structured manner through database servers[1]. The advantage of using SQL is that it can handle the large amount of data. The need for the SQL here is to run the aggregate queries on GCP to check for the accurate results obtained from the query. The SQL queries for the State Government Employee with a Bachelor's degree is considered as an example to explain the method to write SQL Queries:

1. **SELECT** statement is used to select the data from the dataset and display the result in the form of a result table. When we need to select the attributes for which the query has to produce the result, we make use of **SELECT** statement . The example for the **SELECT** statement is given below which selects lower age, upper age, Gender, Zip code, Employment and Education.

```
SELECT
  Lower_Age,
  Upper_Age,
  Gender,
  Zip_Code,
  Education,
  Employment
```

Fig.3.3: **SELECT** Statement

2. **FROM** command here specifies from which table we have to select data or delete data.

```
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Masked_Microdata_1000_Records_k_3_v2`
```

Fig.3.4: **FROM** Command

3. **WHERE** clause here is used to obtain the information of those records which fulfill the specified condition. This clause is used whenever we want to look for the particular attribute or for some condition. The **WHERE** clause in the fig.3.5 choose the records of the people with education as Bachelor's and a State Government Employee.

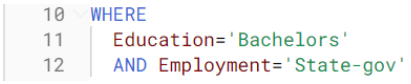


Fig.3.5: **WHERE** clause

4. **ORDER BY** keyword here is used to arrange or sort the data in ascending or descending order. By default, the **ORDER BY** keyword sorts the data in the ascending order. The **ORDER BY** keyword is used whenever there is a requirement of arranging the data based on the attribute or condition mentioned. **ORDER BY** keyword in the fig.3.6 sorts the data based on the GROUP ID.

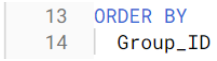


Fig.3.6 : **ORDER BY** keyword of the SQL Query

5. **LIMIT** keyword specifies the number of allowable errors . The **LIMIT** keyword in the fig.3.7 specifies that the number of allowable error is 1000.

Fig.3.7: **LIMIT** keyword of the SQL Query

The method for writing the SQL Query mentioned above is followed throughout the work. The Query selects the attributes from the microdata data , looks for the specified condition or case ,sorts the data either in ascending or descending order and displays the result in the tabular form for the number of records specified.

The example of the Query and its associated Result is shown below in fig.3.8.

16 lines (16 sloc) | 271 Bytes Raw Blame

```
SELECT
  Lower_Age,
  Upper_Age,
  Gender,
  Zip_Code,
  Education,
  Employment
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Masked_Microdata_1000_Records_k_3_v2`
WHERE
  Education='Bachelors'
  AND Employment='State-gov'
ORDER BY
  Group_ID
LIMIT
  1000
```

Fig.3.8: Query for Education=Bachelors and Employment= State- Government Employee

Age	Gender	Zip Code	Education	Employment
(30 - 39)	M/F	77***	Bachelors	State-gov
(30 - 39)	M/F	141***	Bachelors	State-gov
(20 - 49)	M/F	267***	Bachelors	State-gov
(20 - 39)	M/F	149***	Bachelors	State-gov
(30 - 59)	M/F	98***	Bachelors	State-gov
(50 - 79)	M/F	288***	Bachelors	State-gov
(30 - 59)	M/F	92***	Bachelors	State-gov
(20 - 29)	M/F	70***	Bachelors	State-gov
(30 - 39)	M/F	169***	Bachelors	State-gov
(20 - 29)	M/F	335***	Bachelors	State-gov
(40 - 69)	M/F	260***	Bachelors	State-gov
(40 - 49)	M/F	193***	Bachelors	State-gov

Fig.3.9: Result Obtained from the Query shown in the Fig. 3.8

The Query has selected all the records which satisfies the condition of Education=Bachelors and Employment=State- Government from all the records present in the microdata table. The Queries that have been run for the different conditions follow the same method discussed above, is explained in detail with their results and comparison in the result section.

RESULT

The Incremental Diversity algorithm was implemented on a dataset containing 5000 records, which includes both quasi-identifiers and sensitive attributes. The original microdata table is shown in Fig. 3.1. After implementing the algorithm, the masked microdata table is shown in Fig. 3.2. The results of the algorithm are compared to the original microdata table to evaluate the effectiveness of the algorithm in protecting the privacy of the data. To test the accuracy of the results obtained from the Incremental Diversity algorithm, aggregate SQL queries were run on the Google Cloud Platform (GCP). The method for writing the SQL queries is explained in detail in the methodology section. The queries were run for different conditions, and the results are compared in this section.

An example query and its associated result is shown in Fig. 3.8 and Fig. 3.9, respectively. The query selects all records that satisfy the condition of Education=Bachelors and Employment=State-Government from all records present in the microdata table. The query returns a tabular form of the result, which shows the number of records that match the specified conditions. Other queries were run for different conditions, such as employment status, education level, and age range. The results of these queries are shown in tables and figures throughout this section. The results demonstrate that the Incremental Diversity algorithm effectively protects the privacy of the data while still allowing for useful data analysis.

A comparison of the original microdata table and the masked microdata table after implementing the Incremental Diversity algorithm is also presented in this section. The comparison shows that the algorithm reduces the number of residue records and prevents significant data loss. The comparison also shows that the algorithm does not have a significant impact on the accuracy of the data analysis.

Overall, the results of this study demonstrate that the proposed Incremental Diversity algorithm is an effective method for protecting the privacy of sensitive data while still allowing for useful data analysis. The algorithm can be used in various industries and applications where privacy is a concern.

QUERY RESULTS

Query 1: State – gov Bachelors Details

```
SELECT
  Lower_Age,
  Upper_Age,
  Gender,
  Zip_Code,
  Education,
  Employment
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Masked_Microdata_1000_Records_k_3_v2`
WHERE
  Education='Bachelors'
  AND Employment='State-gov'
ORDER BY
  Group_ID
LIMIT
  1000
```

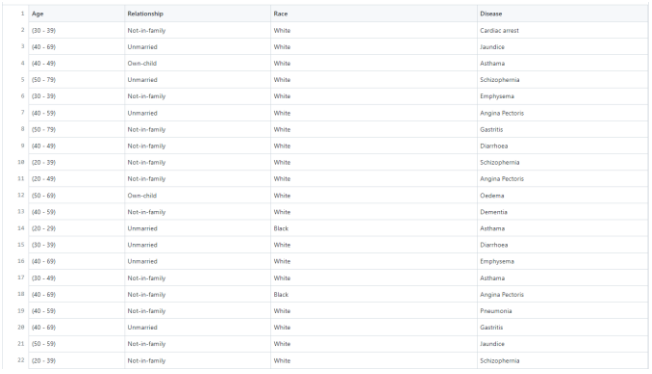
	Age	Gender	Zip_Code	Education	Employment
1	(30 - 39)	M/F	77***	Bachelors	State-gov
2	(30 - 39)	M/F	141***	Bachelors	State-gov
3	(20 - 49)	M/F	267***	Bachelors	State-gov
4	(20 - 39)	M/F	149***	Bachelors	State-gov
5	(30 - 59)	M/F	98***	Bachelors	State-gov
6	(50 - 79)	M/F	288***	Bachelors	State-gov
7	(30 - 59)	M/F	92***	Bachelors	State-gov
8	(20 - 29)	M/F	70***	Bachelors	State-gov
9	(30 - 39)	M/F	169***	Bachelors	State-gov
10	(20 - 29)	M/F	335***	Bachelors	State-gov
11	(40 - 69)	M/F	260***	Bachelors	State-gov
12	(40 - 49)	M/F	193***	Bachelors	State-gov
13	(20 - 39)	M/F	356***	Bachelors	State-gov

Fig 1: Masked Table for State – gov Bachelors Query

In this query, we have asked the program to show age, gender, zip code, education and employment from the Masked Table in the results. In this, we have selected 1000 records in which Education is ‘Bachelors’ and Employment is ‘State-gov’ and have ordered them based on Group ID. Age and gender have been generalized and zip code is masked.

Query 2: Divorcee

```
SELECT
  Lower_Age,
  Upper_Age,
  Relationship,
  Race,
  Disease
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Masked_Microdata_1000_Records_k_3_v2`
WHERE
  Marital_Status='Divorced'
ORDER BY
  Group_ID
LIMIT
  1000
```



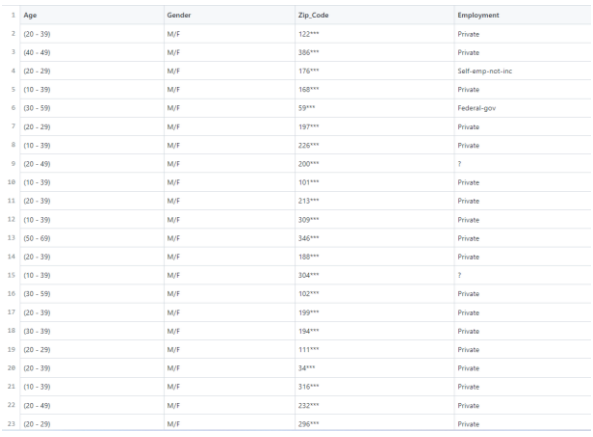
	Age	Relationship	Race	Disease
1	(30 - 39)	Not-in-family	White	Cardiac arrest
2	(40 - 69)	Unmarried	White	Jaundice
3	(40 - 49)	Own-child	White	Asthama
4	(50 - 79)	Unmarried	White	Schizophemia
5	(30 - 39)	Not-in-family	White	Emphysema
6	(40 - 59)	Unmarried	White	Angina Pectoris
7	(50 - 79)	Not-in-family	White	Gastritis
8	(40 - 49)	Not-in-family	White	Diarrhoea
9	(20 - 39)	Not-in-family	White	Schizophemia
10	(20 - 49)	Not-in-family	White	Angina Pectoris
11	(50 - 69)	Own-child	White	Oedema
12	(40 - 59)	Not-in-family	White	Dementia
13	(20 - 29)	Unmarried	Black	Asthama
14	(30 - 39)	Unmarried	White	Diarrhoea
15	(40 - 69)	Unmarried	White	Emphysema
16	(30 - 49)	Not-in-family	White	Asthama
17	(40 - 69)	Not-in-family	Black	Angina Pectoris
18	(40 - 59)	Not-in-family	White	Pneumonia
19	(40 - 69)	Unmarried	White	Gastritis
20	(50 - 59)	Not-in-family	White	Jaundice
21	(20 - 39)	Not-in-family	White	Schizophemia

Fig 2: Masked Table for Divorcee Query

In this query, we are displaying the Age, Relationship, Race and Disease Values from the Masked Table. The program is asked to display a maximum of 1000 records from the Masked Table which have been ordered by Group Id and having Marital Status as ‘Divorced’. From the result, we can observe that most of the people don’t have any family and majority of them are white.

Query 3: Own-Child-White

```
SELECT
  Lower_Age,
  Upper_Age,
  Gender,
  Zip_Code,
  Employment
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Masked_Microdata_1000_Records_k_3_v2`
WHERE
  Relationship='Own-child'
  AND Race='White'
ORDER BY
  Group_ID
LIMIT
  1000
```



	Age	Gender	Zip_Code	Employment
1	(20 - 39)	M/F	122***	Private
2	(40 - 49)	M/F	386***	Private
3	(20 - 29)	M/F	176***	Self-emp-not-inc
4	(10 - 39)	M/F	168***	Private
5	(30 - 59)	M/F	59***	Federal-gov
6	(20 - 29)	M/F	197***	Private
7	(10 - 39)	M/F	226***	Private
8	(20 - 49)	M/F	200***	?
9	(10 - 39)	M/F	101***	Private
10	(20 - 39)	M/F	213***	Private
11	(10 - 39)	M/F	309***	Private
12	(50 - 69)	M/F	346***	Private
13	(20 - 39)	M/F	188***	Private
14	(10 - 39)	M/F	304***	?
15	(30 - 59)	M/F	102***	Private
16	(20 - 39)	M/F	199***	Private
17	(30 - 39)	M/F	194***	Private
18	(20 - 29)	M/F	111***	Private
19	(20 - 39)	M/F	34***	Private
20	(10 - 39)	M/F	316***	Private
21	(20 - 49)	M/F	232***	Private
22	(20 - 29)	M/F	296***	Private

Fig 3: Masked Table for Own-Child-White Query

The Age, Gender, Zip Code and Employment of the people who are white in color and having relationship as ‘Own-child’ is shown from the Masked Table. A maximum of 1000 records ordered based on Group ID and satisfying the given conditions are shown. From the result we can see that most of the people lie in the 20 – 40 age and have Private employment.

Query 4: Widowed-Salary>50k

```
SELECT
  Lower_Age,
  Upper_Age,
  Gender,
  Zip_Code,
  Education
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Masked_Microdata_1000_Records_k_3_v2`
WHERE
  Marital_Status='Widowed'
  AND Salary='>50K'
ORDER BY
  Group_ID
LIMIT
  1000
```

Age	Gender	Zip_Code	Education
(50 - 79)	M/F	34***	HS-grad
(60 - 79)	M/F	99***	Masters
(60 - 69)	M/F	105***	Bachelors
(40 - 59)	M/F	318***	HS-grad

Fig 4: Masked Table for Widowed-Salary>50k Query

In this query, the Age, Gender, Zip Code and Education records of the people who are widow and having a salary greater than 50k will be shown from the Masked Table. The records obtained by satisfying the given conditions have been ordered based on their Group ID. From the table, we can observe that the condition of salary greater than 50k is met by those people who have very high Education.

Analysis of the Queries

To prove the accuracy of the Incremental diversity algorithm we compare the results obtained from the queries for original microdata table and masked microdata table.

1. Comparison of Results obtained from the Original Dataset Query and the Masked Microdata Query for Age.

```
SELECT
  Age,
  Gender,
  Zip_Code,
  Education,
  Employment
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Original_Microdata_1000_Records_k_3`
WHERE
  Age=35
LIMIT
  1000
```

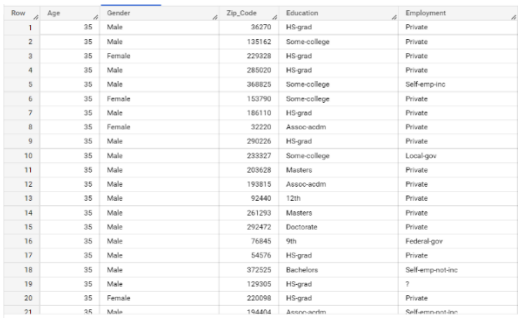
Fig 1: Age Query for Original Microdata Table

```
SELECT
  Lower_Age,
  Upper_Age,
  Gender,
  Zip_Code,
  Education,
  Employment
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Masked_Microdata_1000_Records_k_3_v2`
WHERE
  Lower_Age>=30
  AND Upper_Age<=40
ORDER BY
  Group_ID
LIMIT
  1000
```

Fig 2: Age Query for Masked Microdata Table

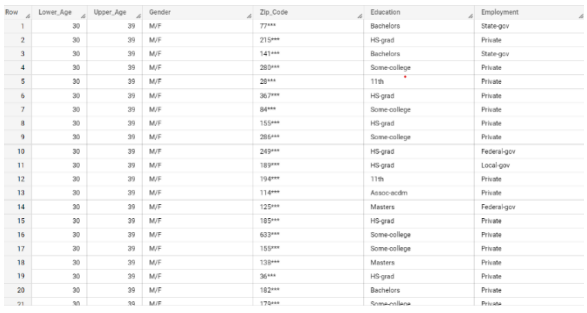
The Comparison of the Results obtained from the Original Dataset Query and the Masked Microdata Query (Age is Generalized) is done here. The Comparison reveals that the number of records in the Original Dataset is significantly lower than the number of records obtained from the Masked Microdata because more generalizations are made there. In Fig. 3, the query

executed for the Original Dataset where Age = 35 showed only 28 records, whereas the query executed for the Masked Microdata where Age is generalized between 30 and 40 showed 80 records (Fig. 4), which is much more than the Original Dataset. The Generalization of Age has increased the number of records in order to preserve the privacy of data[8].



Row	Age	Gender	Zip_Code	Education	Employment
1	35	Male	36370	HS-grad	Private
2	35	Male	135162	Some-college	Private
3	35	Female	229328	HS-grad	Private
4	35	Male	285020	HS-grad	Private
5	35	Male	368825	Some-college	Self-emp-inc
6	35	Female	153790	Some-college	Private
7	35	Male	186110	HS-grad	Private
8	35	Female	32220	Assoc-acdm	Private
9	35	Male	290226	HS-grad	Private
10	35	Male	233327	Some-college	Local-gov
11	35	Male	203628	Masters	Private
12	35	Male	193815	Assoc-acdm	Private
13	35	Male	92440	12th	Private
14	35	Male	261293	Masters	Private
15	35	Male	292472	Doctorate	Private
16	35	Male	76845	9th	Federal-gov
17	35	Male	54576	HS-grad	Private
18	35	Male	372525	Bachelors	Self-emp-not-inc
19	35	Male	129305	HS-grad	?
20	35	Female	220098	HS-grad	Private
21	35	Male	194494	Assoc-acdm	Self-emp-not-inc

Fig 3: Original Microdata Table for Age



Row	Lower_Age	Upper_Age	Gender	Zip_Code	Education	Employment
1	30	39	M/F	77***	Bachelors	State-gov
2	30	39	M/F	215***	HS-grad	Private
3	30	39	M/F	141***	Bachelors	State-gov
4	30	39	M/F	280***	Some-college	Private
5	30	39	M/F	28***	11th	Private
6	30	39	M/F	367***	HS-grad	Private
7	30	39	M/F	84***	Some-college	Private
8	30	39	M/F	159***	HS-grad	Private
9	30	39	M/F	286***	Some-college	Private
10	30	39	M/F	249***	HS-grad	Federal-gov
11	30	39	M/F	189***	HS-grad	Local-gov
12	30	39	M/F	194***	11th	Private
13	30	39	M/F	114***	Assoc-acdm	Private
14	30	39	M/F	125***	Masters	Federal-gov
15	30	39	M/F	185***	HS-grad	Private
16	30	39	M/F	603***	Some-college	Private
17	30	39	M/F	159***	Some-college	Private
18	30	39	M/F	138***	Masters	Private
19	30	39	M/F	36***	HS-grad	Private
20	30	39	M/F	182***	Bachelors	Private
21	30	39	M/F	178***	Some-college	Private

Fig 4: Masked Microdata Table for Age

2. Comparison of Results obtained from the Original Dataset Query and the Masked Microdata Query for Gender.

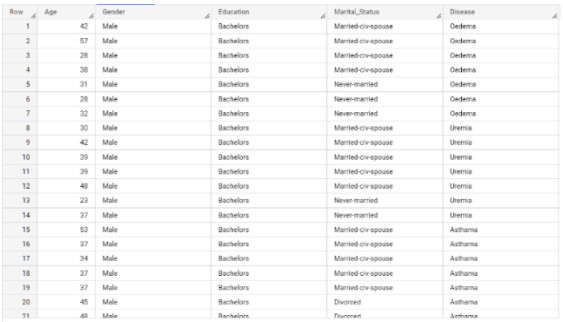
```
SELECT
  Age,
  Gender,
  Education,
  Marital_Status,
  Disease
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Original_Microdata_1000_Records_k_3`
WHERE
  Gender="Male"
  AND
  Education="Bachelors"
LIMIT
  1000
```

Fig 5: Gender Query for Original Microdata Table

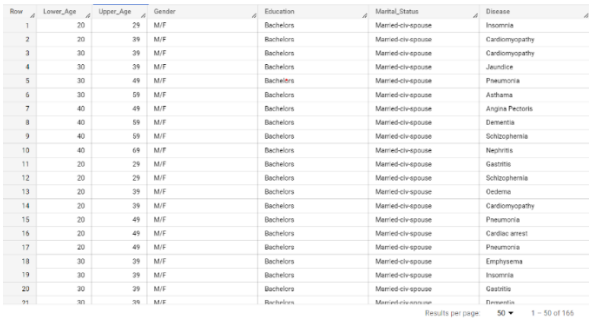
```
SELECT
  Lower_Age,
  Upper_Age,
  Gender,
  Education,
  Marital_Status,
  Disease
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Masked_Microdata_1000_Records_k_3_v2`
WHERE
  Education="Bachelors"
LIMIT
  1000
```

Fig 6: Gender Query for Masked Microdata Table

The Result of the Query for the Original Dataset consisting of Age, Gender, Marital Status, Education, Disease is compared with the result of the Query for Masked Microdata consisting of Upper Age, Lower Age (Age is generalized in this case as shown in Fig 8.), Gender, Marital Status, Education, Disease. The Comparison shows that the Original Dataset has less records as query results, Masked Microdata has more records due to more generalization done. In Fig7., the query executed for the Original Dataset where Gender="Male “and Education="Bachelors “showed 117 records whereas, the query that has been executed for the Masked Microdata with gender generalized to M/F has 166 records (Fig 8.) which is more than the records of the Original Dataset. Generalization has increased the number of records in order to preserve the privacy of the data[7].



Row	Age	Gender	Education	Marital_Status	Disease
1	42	Male	Bachelors	Married-civ-spouse	Oedema
2	57	Male	Bachelors	Married-civ-spouse	Oedema
3	28	Male	Bachelors	Married-civ-spouse	Oedema
4	38	Male	Bachelors	Married-civ-spouse	Oedema
5	31	Male	Bachelors	Never-married	Oedema
6	28	Male	Bachelors	Never-married	Oedema
7	32	Male	Bachelors	Never-married	Oedema
8	30	Male	Bachelors	Married-civ-spouse	Uremia
9	42	Male	Bachelors	Married-civ-spouse	Uremia
10	39	Male	Bachelors	Married-civ-spouse	Uremia
11	39	Male	Bachelors	Married-civ-spouse	Uremia
12	48	Male	Bachelors	Married-civ-spouse	Uremia
13	23	Male	Bachelors	Never-married	Uremia
14	37	Male	Bachelors	Never-married	Uremia
15	53	Male	Bachelors	Married-civ-spouse	Asthama
16	37	Male	Bachelors	Married-civ-spouse	Asthama
17	34	Male	Bachelors	Married-civ-spouse	Asthama
18	37	Male	Bachelors	Married-civ-spouse	Asthama
19	37	Male	Bachelors	Married-civ-spouse	Asthama
20	45	Male	Bachelors	Divorced	Asthama



Row	Lower_Age	Upper_Age	Gender	Education	Marital_Status	Disease
1	20	29	M/F	Bachelors	Married-civ-spouse	Insomnia
2	20	39	M/F	Bachelors	Married-civ-spouse	Cardiomyopathy
3	30	39	M/F	Bachelors	Married-civ-spouse	Cardiomyopathy
4	30	39	M/F	Bachelors	Married-civ-spouse	Jaundice
5	30	49	M/F	Bachelors	Married-civ-spouse	Pneumonia
6	30	59	M/F	Bachelors	Married-civ-spouse	Asthama
7	40	49	M/F	Bachelors	Married-civ-spouse	Angina Pectoris
8	40	59	M/F	Bachelors	Married-civ-spouse	Dementia
9	40	59	M/F	Bachelors	Married-civ-spouse	Schizophernia
10	40	69	M/F	Bachelors	Married-civ-spouse	Nephritis
11	20	29	M/F	Bachelors	Married-civ-spouse	Gastritis
12	20	29	M/F	Bachelors	Married-civ-spouse	Schizophernia
13	20	39	M/F	Bachelors	Married-civ-spouse	Oedema
14	20	39	M/F	Bachelors	Married-civ-spouse	Cardiomyopathy
15	20	49	M/F	Bachelors	Married-civ-spouse	Pneumonia
16	20	49	M/F	Bachelors	Married-civ-spouse	Cardiac arrest
17	20	49	M/F	Bachelors	Married-civ-spouse	Pneumonia
18	30	39	M/F	Bachelors	Married-civ-spouse	Emphysema
19	30	39	M/F	Bachelors	Married-civ-spouse	Insomnia
20	30	29	M/F	Bachelors	Married-civ-spouse	Gastritis

Results per page: 50 1 – 50 of 166

Fig 7: Original Microdata Table for Gender

Fig 8: Masked Microdata Table for Gender

3. Comparison of Results obtained from the Original Dataset Query and the Masked Microdata Query for Zip code.

```
SELECT
  Age,
  Gender,
  Zip_Code,
  Employment,
  Marital_Status
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Original_Microdata_1000_Records_k_3`
WHERE
  Zip_Code=51618
LIMIT
  1000
```

Fig 9: Zip Code Query for Original Microdata Table

```
SELECT
  Lower_Age,
  Upper_Age,
  Gender,
  Zip_Code,
  Employment,
  Marital_Status
FROM
  `plenary-chalice-369413.Incremental_Diversity_Dataset.Masked_Microdata_1000_Records_k_3_v2`
WHERE
  Zip_Code LIKE "51%"
LIMIT
  1000
```

Fig 10: Zip Code Query for Masked Microdata Table

Here, comparison of query results for Original Dataset and Masked Microdata is done with respect to zip code. Similar to the comparison of results of queries with respect to gender and age , the number of records in Original Dataset is less than the number of records in Masked Microdata. In Fig. 11, the query executed for the Original Dataset where Zip Code = 51618 showed only 1 record whereas the query executed for the Masked Microdata where Zip Code is generalized as 51*** showed 9 records (Fig 12)[5]. This is because of generalization in Masked Microdata.

Row	Age	Gender	Zip_Code	Employment	Marital_Status
1	46	Female	51618	Private	Married-civ-spouse

Fig 11: Original Microdata Table for Zip Code

Row	Lower_Age	Upper_Age	Gender	Zip_Code	Employment	Marital_Status
1	40	49	M/F	51***	Private	Married-civ-spouse
2	40	59	M/F	51***	?	Married-civ-spouse
3	40	69	M/F	51***	Private	Married-civ-spouse
4	50	79	M/F	51***	Private	Married-civ-spouse
5	50	79	M/F	51***	Self-emp-inc	Married-civ-spouse
6	30	39	M/F	51***	Private	Divorced
7	30	49	M/F	51***	Private	Divorced
8	60	79	M/F	51***	Private	Divorced
9	90	119	M/F	51***	Private	Never-married

Fig 12: Original Microdata Table for Zip Code

The results obtained through the comparison of age, gender and zip code reveals that the number of records obtained from the original microdata table is less than the number of records obtained from the masked microdata table. Since, the number of records in the masked microdata table is more, it ensures security and privacy as more number of records has been generalized or protected from the background attacks[3]. This proves the efficiency and accuracy of the Incremental Diversity algorithm in protecting the privacy of the published data[4]. This model outstands the other anonymity models in providing very a smaller number of record loss and more accurate results.

REFERENCES:

1.SQL:From traditional databases to the Big data :Yasin N Silva, Isadora Almeida, Michell Queiroz published in 47th ACM technical Symposium.

2.Estimating progress of Long running SQL queries : Surajit Chaudary, Vivek R Narasayya, Ravishankar Ramamurthy published on International Conference on Management of Data, Paris, France.

3. Privacy-preserving verification of aggregate queries on outsourced databases: Stuart Haber, William G. Horne, Tomas Sander, Danfeng Yao, University of South Florida published on January 2007.

4. Privacy-Preserving Computation and Verification of Aggregate Queries on Outsourced Databases : Brian Thompson , Stuart Haber, William G. Horne , Tomas Sander , and Danfeng Yao1.

5.Authenticating Aggregate Range Queries over Multidimensional Dataset: Jia Xu, Ee-Chien Chang, National University of Singapore Department of Computer Science.

6.Authenticating Aggregate Queries over Set-Valued Data with Confidentiality: Cheng Xu , Qian Chen , Haibo Hu , Jianliang Xu , Xiaojun Hei, Department of Computer Science, Hong Kong Baptist University, Hong Kong.

7.Answering Aggregation Queries in a Secure System Model:January 2007,Conference: Proceedings of the 33rd International Conference on Very Large Databases, University of Vienna, Austria, September 23-27, 2007, Ting Jian Ge:University of Massachusetts, Lowell Stan, Zdonik.

8.Efficient Table Anonymization for Aggregate Query Answering: C Procopiuc, D.Srivatsava, Published on March 29,2009 on International Conference on Data Engineering.