RV Educational Institutions ®
RV College of Engineering ®

Autonomous Institution Affiliated to Visvesvaraya Technological University, Belagavi

Approved by AICTE, New Delhi

DATA ANONYMIZATION TECHNIQUES FOR MITIGATION OF PRIVACY ATTACKS IN PRIVACY PRESERVING DATA PUBLISHING

-Hardik Pawar, Tanmay S Lal, S.M. Ashiq, M Raza

Internship coordinator : Prof. Veena Gadad (Asst. Professor, Dept. of CSE)
Dr. Sowmyarani C N (Assoc. Professor, Dept. of CSE)

R.V. College of Engineering, Bangalore-59

1) ABOUT THE COMPANY / THE COE

Research Projects:

1. Developing a novel algorithm for privacy preserving data publishing multiple sensitive attributes
2. Validating the anonymized data using aggregate queries in public cloud

2) TOOLS AND TECHNOLOGIES USED

SOFTWARE COMPONENTS	SPECIFICATIONS
Python (time, copy, pandas libraries)	3.9
Google Cloud Platform (GCP)	
Git & GitHub (Version Control)	
Web Development	HTML, CSS, JavaScript, React.js

3) PROBLEM STATEMENT

Privacy is the state of being free from public scrutiny or from having your secrets or personal information shared. Data protection is important, since it prevents the information of an organization from fraudulent activities, hacking, phishing, and identity theft.

The issue with the previous privacy preserving algorithms was that it was either lacking the multiple sensitive attribute or loss of information.

4) OBJECTIVE OF INTERNSHIP

While publishing data, utmost care should be taken that the data should be published in such a manner that it ensures privacy.
So, our aim is to publish data by keeping in balance the privacy and its usefulness for researchers in their research by analyzation of the data.
The Research Project deals with anonymizing the given microdata table based on the incremental diversity algorithm and publishing the mass dataset on GCP.
The output of our project contains various anonymity techniques like k-anonymity , l-diversity , anatomy ,l-e diversity etc.

5) METHODOLOGY

1. The microdata table is imported in a standardized manner and stored with the implementation of a nested dictionary.
2. The group id or more specifically the equivalence class number of a particular record is calculated with the formula given below. $Group\ ID$= $\lfloor (n-1)/k \rfloor + 1, n \in [1,N]$
3. Diversification of the records via Incremental Diversity.
4. The various performance parameters to evaluate and analyse the algorithm such as Code Runtime, Residue Percentage, Diversity Percentage

6) RESULTS OBTAINED

The success of our work is that it outstands other researches is that for the maximum number of records the amount of information loss is very less.

Incremental diversity outperforms l, e diversity in terms of faster time performance (code runtime) and overall decrease in residue records percentage. .

7) APPLICATIONS

The Incremental Diversity algorithm can be particularly useful when we want to produce a masked dataset with lesser no. of residue records, finding application in civil datasets that require the data of all people to be present, allowing for a loss in a minimal amount of diversity.

8) CONCLUSION

l, e diversity algorithm employs the use of a primary sensitive attribute with lesser no. of parents in the semantic hierarchical tree (Eg. Marital Status). The records in the equivalence class are diversified such that only one common parent for the sensitive attribute should be present in a particular equivalence class. Incremental diversity algorithm makes use of a primary sensitive attribute with more no, of parents in the semantic hierarchical tree (Eg. Disease). It follows the same condition as l, e diversity where only non-repeating parents should be present in each equivalence class. In addition, it also performs the incremental diversification for secondary, tertiary and quaternary sensitive attributes thereby increasing the diversity characteristics for the equivalence classes and also the entire table.

9) TAKE AWAY

- We as a team had hands-on experience in many privacy preserving algorithms and also explored the cloud architecture of GCP.
- We familiarized ourselves with Buckets, Queries and a lot of Web Development in the course of the Research Project.
- We developed an interactive website to display the privacy preserving algorithm "Incremental Diversity" that we had created.
- And compared the results of running various permutations and combinations of queries on the masked and original datasets and then comparing the results.

10) REFERENCES

K-anonymity: A model for protecting the privacy: LATANYA SWEENEY, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.
(1, m, d) - Anonymity: A Resisting Similarity Attack Model for Multiple Sensitive Attributes: Junjie Jia, Luting Chen School of Computer Science and Engineering, Northwest Normal University Lanzhou, China.
Simple Distribution of Sensitive Values for Multiple Sensitive Attributes in Privacy Preserving Data Publishing to Achieve Anatomy: Widodo Informatics Education Universitas Negeri Jakarta, Indonesia, Murein Nugraheni Information Systems and Technology Universitas, Negeri Jakarta, Indonesia. Irma Permata.

TEAM MEMBERS BIOGRAPHY

Hardik Hiraman Pawar (1RV21CS046): 3rd Sem CSE
Proficient in Python, Web Dev (React.js, Node.js, MongoDB), ML, Computer Vision (OpenCV), Java (Android Studio), C, C++. Completed Udemy Courses in various fields such as Web Development, Machine Learning & Deep Learning, Python & C++ Bootcamp, along with Developing a Multithreaded Kernel in Linux. I maintain a GitHub Repository with 50+ projects ranging from LeetCode and CodeChef solutions to Web Dev, ML & Computer Vision Projects.

Tanmay S Lal (1RV21CS176): 3rd Sem CSE
Proficient in Python, Web Dev, Android Dev, Java, C, Dart Language, Flutter

S Mohammed Ashiq (1RV21CS132): 3rd Sem CSE
Proficient in Python, Web Dev, Android Dev, Java, C, Dart Language, Flutter

Mohammed Raza (1RV21CS093): 3rd Sem CSE
Proficient in Web Dev, C