Privacy Preserving Data Mining

Minor project II

Submitted by:  
Dharmesh Malav (19803005)  
Nikhil Paleti (19803024)  
Divyanshu Tiwari (19803025)

Under the supervision of:  
Mrs. Niyati Aggrawal



Department of CSE/IT  
Jaypee Institute of Information Technology University, Noida

May 2022

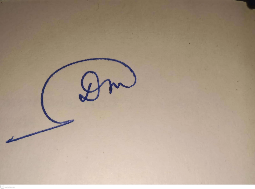
ACKNOWLEDGEMENT

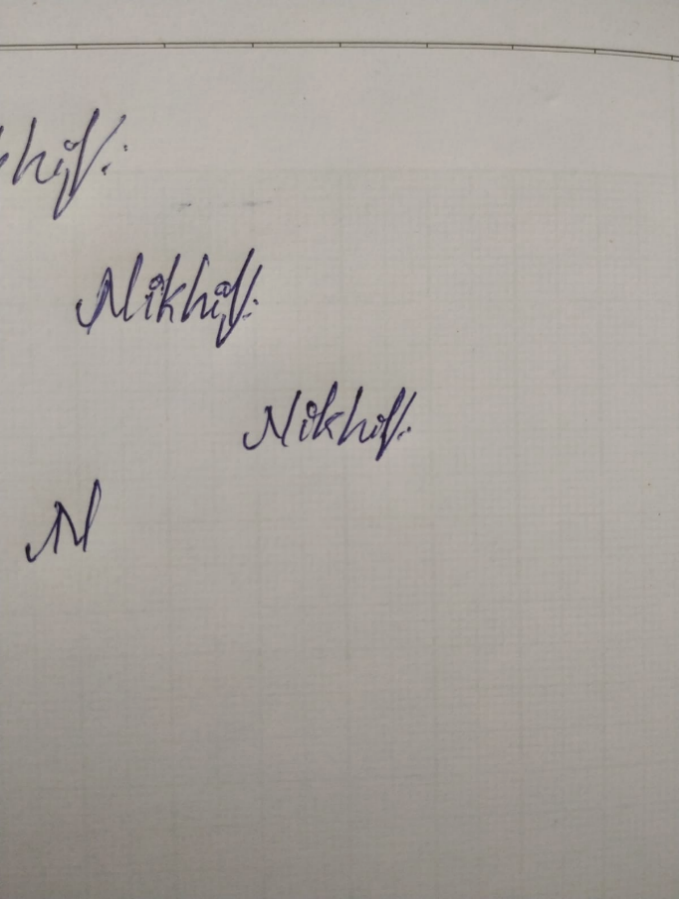
I would like to place on record my deep sense of gratitude to Mrs. Niyati Aggrawal , Designation, Jaypee Institute of Information Technology, India for his/her generous guidance, help and useful suggestions.

I express my sincere gratitude to Mrs. Niyati Aggrawal , Dept. of CSE, India, for his/her stimulating guidance, continuous encouragement and supervision throughout the course of present work.

I also wish to extend my thanks to \_\_\_\_\_\_\_\_\_\_\_\_\_ and other classmates for their insightful comments and constructive suggestions to improve the quality of this project work.

Signature(s) of Students

Dharmesh Malav (19803005) 

Nikhil Paleti (19803024)

Divyanshu Tiwari (19803025) 

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and beliefs, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma from a university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: Jaypee Institute of information Technology, Noida

Date: 16 May, 2022

Name: Dharmesh Malav

Enrolment No.: 19803005

Name: Nikhil Paleti

Enrolment No.: 19803024

Name: Divyanshu Tiwari

Enrolment No.:19803025

CERTIFICATE

This is to certify that the work titled “Privacy Preserving Data Mining” submitted by Nikhil Paleti , Dharmesh Malav and Divyanshu Tiwari of B.Tech of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any other degree or diploma.

Digital Signature of Supervisor

Name of Supervisor : Mrs. Niyati Aggrawal

Designation :

Date :

ABSTRACT

The objective of the proposed project is to look into and investigate about Privacy Preserving Techniques that are existing, and employed in the real world. The proposed project would also apply, analyze and evaluate the privacy preserving techniques as deployed, on a real-life dataset.

TABLE OF CONTENTS

Page No.

*Acknowledgement i*

*Declaration ii*

*Certificate iii*

*Abstract iv*

*Table of Contents* *v*

*List of Figures* *vi*

Chapter 1: INTRODUCTION 1

Chapter 2: BACKGROUND STUDY 2

References

LIST OF FIGURES

Figure Title Page

3.1 ........................................................................... xx

4.1 Vertical Manifold. ........................................................................................ xx

4.2 20O Bend Manifold......................................................................................... xx

4.3 90O Bend Manifold....................................................................................... xx

4.4 Spiral Manifold .............................................................................................. xx

4.5 Spiral Manifold Configuration (θ = 225o) ..................................................... xx

4.6 Spiral Manifold with Different Flow Entry Angles (20O, 32.5O and 45O) ........ xx

INTRODUCTION

Privacy is the point of conversation of many people lately, because of the complexity, scale and importance of data, more specifically – Personal Data, in the current era of computing, which relies heavily on such data to train Soft Computing models to personalize our experiences, deliver advertisements, and more. This data, however, is often stored as plain text, often unsecured, on databases, to be processed as quickly as required, which is needless to say, a major security and privacy risk for end users, whose data is in question over here.

This calls for research and work to be done on techniques to protect the privacy of the users who have their data spread across the internet, while simultaneously not hindering the usability of the data, or increasing processing overhead, which would make large organizations not consider these.

There are 3 major privacy preserving techniques that have been coined and researched to an extent: k-anonymity, l-diversity and t-closeness

However, the disappointing fact is that Privacy Preservation techniques like these are almost completely unused in the real world, in the field of data mining, even by major organizations like Apple, Tesla or Facebook, who collect user data on a large scale for purposes like training their AI/Soft Computing Models, or more, simply because of the time involved in these, and the lack of perceived monetary benefit.

This project aims to throw more light, and more “HITS” on the topic of privacy, and the (fairly) easy to utilize techniques that can go a long way in protecting consumer privacy, in case of data breaches.

BACKGROUND STUDY

k-Anonymity**:** “K-anonymity is a property of a dataset that indicates the re-identifiability of its records. A dataset is k-anonymous if quasi-identifiers for each person in the dataset are identical to at least k – 1 other people from the dataset.” [3]

K-Anonymity has been proposed as a mechanism for protecting privacy in microdata publishing, and numerous recoding "models" have been considered for achieving k-anonymity. [3]

l-Diversity**:**“A q⋆-block is l-diverse if contains at least l “well-represented” values for the sensitive attribute S. A table is l-diverse if every q⋆-block is l-diverse.” [4]

In l-diversity, we aim to segregate the dataset into “l” diverse sections, where the attributes falling under one of the sections, shares the same combinations of the key attributes.

Although the l–diversity principle represents an important step beyond k–anonymity in protecting sensitive attribute disclosures, it still has some shortcomings.  
l-diversity is most prone to skewness or attribute disclosure attacks.

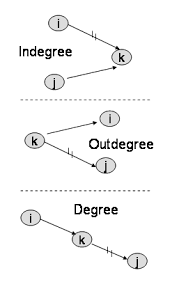
t-closeness**:**“An equivalence class is said to have *t*-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold *t*. A table is said to have *t*-closeness if all equivalence classes have *t*-closeness.” [5]

t-closeness is a privacy model recently defined for data-anonymization. A data set is said to satisfy t-closeness if, for each group of records sharing a combination of key attributes, the Distance between the distribution of a confidential attribute in the group and the distribution of the attribute in the entire data set is no more than a threshold t.

HITS Algorithm**:** Hyperlink-Induced Topic Search (also known as hubs and authorities) is a link analysis algorithm that rates Web pages.   
Despite being originally designed for web pages, this algorithm can be applied in modified form, on any directed graph, including our dataset.

The HITS Algorithm is based on a fundamental model of the internet, which was basically a directional graph. The HITS Algorithm is an iterative algorithm, which works by rating each webpage (Node) according to two metrics

Authority Value, which estimates the value of the content/node itself (in-degree). Hub Value, which estimates the value of its links to other nodes (out-degree) [1].



REQUIREMENT ANALYSIS

Python Packages required:

* Numpy: Numpy is a Python library that provides powerful mathematic manipulation techniques and data structures that are also faster than native Python.
* Pandas: Pandas is a Python library bringing and providing functionality for high-performance data manipulation and analysis tool using its powerful data structures, into Python.
* Matplotlib: Matplotlib was originally created as a MATLAB port on Python, and it provides MATLAB-like functions to plot graphs in Python.

Dataset used/s:

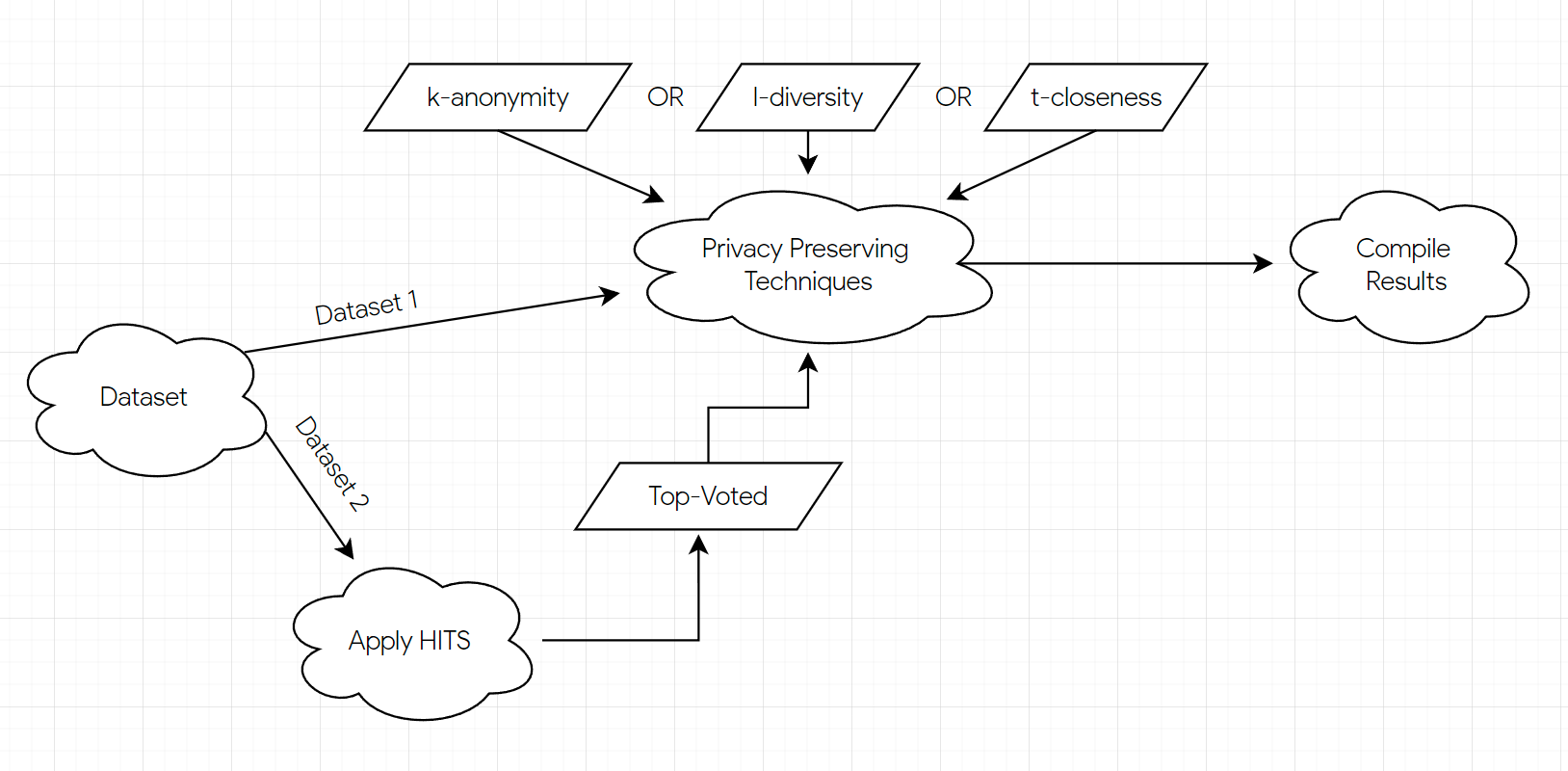
We are using 2 Datasets –

* Human Resources Data Set – Kaggle  
  <https://www.kaggle.com/datasets/rhuebner/human-resources-data-set>   
  This is a publicly available dataset
* A custom, pseudorandom dataset was created, taking inspiration from Wiki\_Votes from Standford SNAP networks (<https://snap.stanford.edu/data/wiki-Vote.html>)  
  We opted to pick a number of “voters” based on their EmpID, who are currently marked “Active” in the company, and made a code to create a random distribution of votes from one employee, to another. They were stored, comma-separated, in a txt file. The left column was voters, and right column is who they voted for.

DETAILED DESIGN

The process of the project can be broken down into the following steps:

1. Data Ingestion and Pre-processing.
2. Apply HITS Algorithm to find the most-voted individuals.
3. Drop entries of least-voted individuals, and apply Privacy Preserving Techniques on the dataset.



As the Project flow suggests, the work-flow begins from procuring, and creating (one of) the two datasets

Then we apply the pre-processing as necessary and store the dataset in a variable, as a pandas dataframe for the first dataset, and the second dataset will be stored as a graph data structure implementation.

The dataset will be used to apply privacy preserving metrics as mentioned above. Before proceeding, however, we apply HITS on our artificially created dataset, to find out the most-voted individuals from the first dataset.

Then the data from both the variables is compared, and results are compiled accordingly

IMPLEMENTATION

Data Preprocessing and Visualization

1. The data from “HRDataset\_v14.csv” in the dataset is ingested into a Pandas Dataframe file.
2. A “graph” data structure is implemented using the custom-created “votes.txt” file.
3. Pandas is used to handle the dataset, and any incomplete data points are removed
4. The outliers will be visualized (scatter plots, or Tensorboard), and removed as necessary.

EXPERIMENTAL RESULTS AND ANALYSIS



Fig 6.1 –

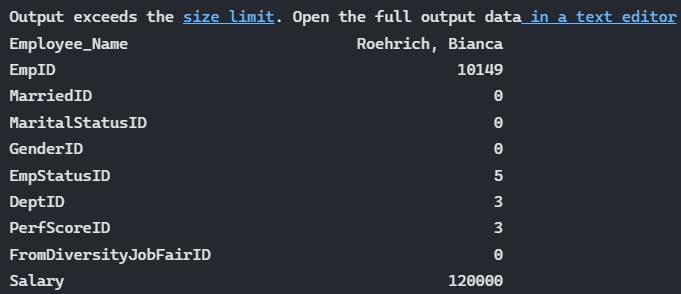


Fig 6.2 –

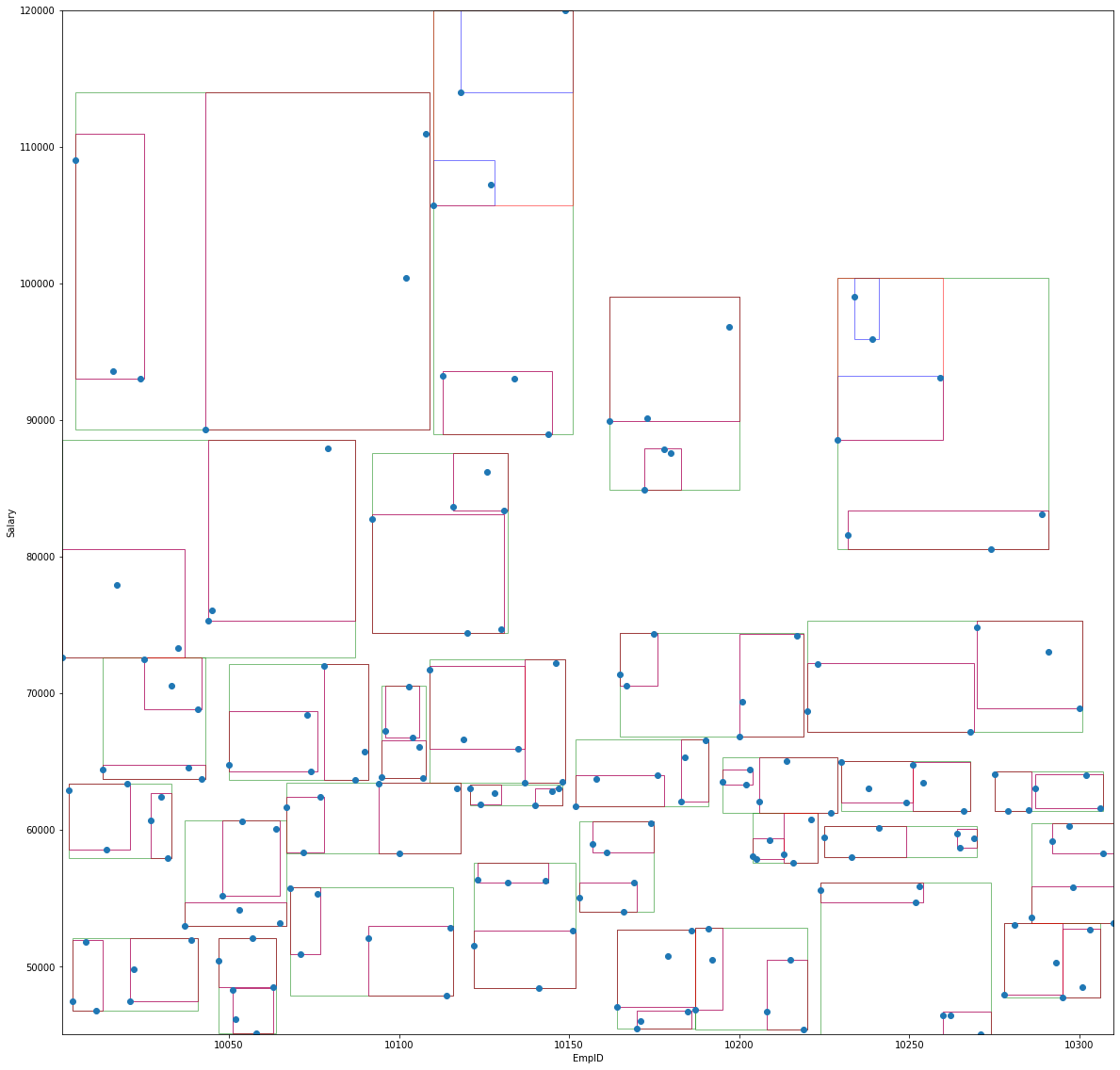


Fig 6.3 –

CONCLUSION OF THE REPORT AND FUTURE SCOPE

This report

REFERENCES

1. Duan, Y., Wang, J., Kam, M., & Canny, J. (2005). Privacy preserving link analysis on dynamic weighted graph. Computational & Mathematical Organization Theory, 11(2), 141-159.
2. Kerschbaum, F., & Schaad, A. (2008, October). Privacy-preserving social network analysis for criminal investigations. In Proceedings of the 7th ACM workshop on Privacy in the electronic society (pp. 9-14)
3. LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006, April). Mondrian multidimensional k-anonymity. In 22nd International conference on data engineering (ICDE'06) (pp. 25-25). IEEE.
4. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 3-es.
5. Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering* (pp. 106-115). IEEE.
6. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.
7. Liu, X., Lin, H., & Zhang, C. (2012). An Improved HITS Algorithm Based on Page-query Similarity and Page Popularity. J. Comput., 7(1), 130-134.
8. <http://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html>