Jaypee Institute of Information Technology, Noida



MINOR PROJECT   
(Synopsis)   
  
Privacy Preserving Data Mining

Team Members:   
Dharmesh Malav - 19803005  
Nikhil Paleti - 19803024  
Divyanshu Tiwari – 19803025

MENTOR – Dr.Niyati Aggrawal

OBJECTIVE

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 evaluate the privacy preserving techniques as deployed, on a real-life dataset.

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 and personalize our experiences.

This data, however, is often stored as plain text, often unsecured, on databases, to be processed as quickly as required.  
This calls for research and work to be done on techniques to protect the privacy of the users who have their data spread on the internet.

There are 3 major privacy preserving techniques that have been coined and researched to an extent.

Privacy Preserving Techniques aim for the property of a dataset, which indicates the re-identifiability of its records

1. k-anonymity: “A dataset is k-anonymous if quasi-identifiers for each person in the dataset are identical to at least k – 1 other people also in the dataset."  
   k-anonymity has 2 major techniques that can be deployed:
   1. Suppression: In this method, certain attributes, or entire columns are replaced by any generic character, like hash ‘#’ or arrow ‘^’. All or some values of a column may be replaced.
   2. Generalization: In this method, individual values of attributes are replaced with a broader category, i.e, a range of values.

k-anonymity is vulnerable to homogeneity attacks

1. 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.”  
   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.

l-diversity is most prone to skewness or attribute disclosure attacks.

1. 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.”

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 webpages, 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).

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, Microsoft or Facebook, who collect user data on a large scale for purposes like training their AT/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

Experimental Design

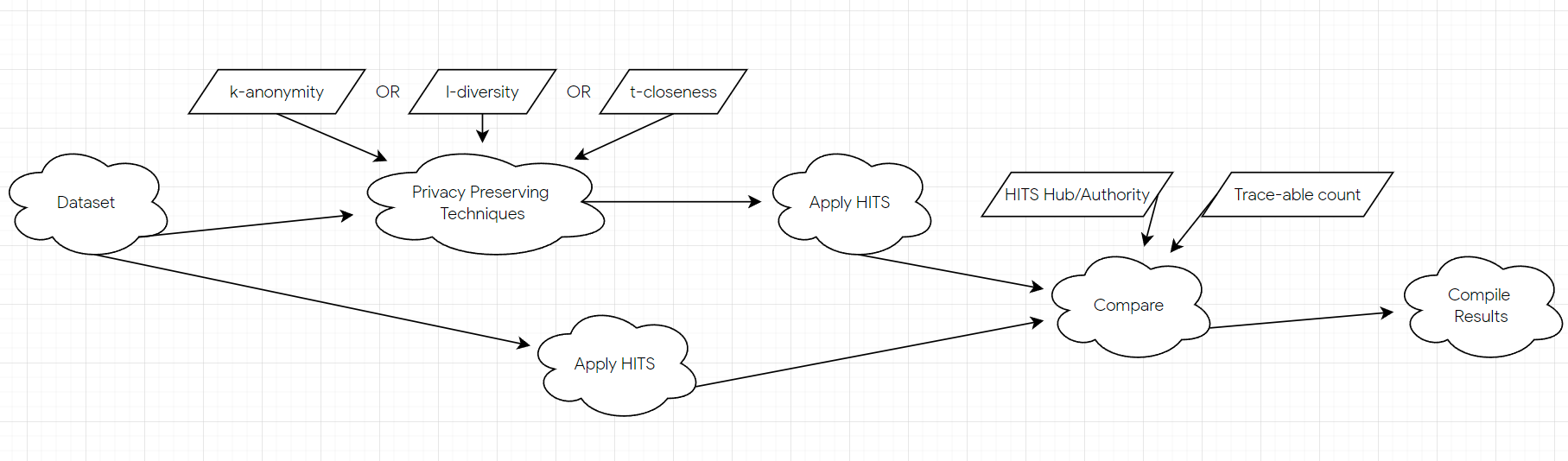
The process of the project can be broken down into the following steps:

1. Data Ingestion and Pre-processing.
2. Apply Privacy Preserving Techniques on the dataset.
3. Apply Performance Measures on Dataset
4. Evaluate scores “before” and “after” the Privacy Preserving Techniques.

Datasets and their Analysis

1. Co-Authorship Network Analysis Dataset  
   <https://www.kaggle.com/code/bkoseoglu/co-authorship-network-analysis/data>   
   This dataset is a freely available dataset, of research papers and scholarly articles published about coronaviruses like CoVID-19 or CoV-2. It contains over 40,000 datapoints about authors, papers published, and collaborations, if any.   
     
   It seems to be an ideal candidate for our research, since the dataset contains certain private information about authors, which is important data that needs to be privatized.

Project Procedure Diagram



As the Project flow suggests, the work-flow begins from procuring a dataset, in our case, the “Co-Authorship Network Analysis Dataset” from Kaggle.   
Then we apply the pre-processing as necessary and store the dataset in a variable, as a numpy array. A graph data structure will also be implemented.   
A copy of the dataset variable will be made and then the copy of the dataset will be used to apply privacy preserving metrics as mentioned above.   
Both of the datasets, before and after privacy-preserving techniques will be run through the HITS Algorithm, to procure Hub/Authority Scores. We also measure count of re-traceable unique data members for the same.   
Then the data from both the variables is compared, and results are compiled accordingly.

Research Methodology

Data Preprocessing and Visualization

1. The data is ingested into a numpy array, from the original 3 CSV files of the dataset.
2. A “graph” data structure is implemented using the author\_papers.csv 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.

Performance Measures

1. **HITS Scores**  
   We will run the HITS Algorithm on the Dataset before and after applying Privacy Preserving Techniques. The delta of the Hub and Authority Scores from HITS would be noted.   
   This measure will be an indicator of the actual usability of the dataset even when the data has been privatized. The delta needs to be minimized to indicate that the effect of the Privacy Preservation is less, or negligible.
2. **Trace-ability.**   
   We will be making count of the number of unique values for authors/papers that we can trace back to, for specific attributes, before and after applying Privacy Preserving Techniques.   
   This count must be reduced after applying privacy measuring techniques since this indicates that the “traceability”, as the name suggests, of the users, is reduced, thus reducing privacy concern

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. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557-570.
4. 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.
5. 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.
6. 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.
7. <http://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html>