DATA ANONYMIZATION

A PROJECT REPORT

Submitted by

ATHUL C K

ERD20CSFI04

to

the APJ Abdul Kalam Technological University

in partial fulfillment of the requirements for the award of the Degree

of

Master of Technology

In

Cyber Forensics and Information Security

DEPT. OF COMPUTER SCIENCE & ENGINEERING

ER&DCI INSTITUTE OF TECHNOLOGY

THIRUVANANTHAPURAM

JULY 2022

DECLARATION

I undersigned hereby declare that the project report “Data Anonymization”, submitted for partial fulfillment of the requirements for the award of degree of Master of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Dr. Dittin Andrews, Ms. Dhanalakshmi M P and Ms. Harsha Gopalakrishnan. This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Athul CK

Place: Thiruvananthapuram

Athul C K

Date: 04-07-2022

DEPT. OF COMPUTER SCIENCE & ENGINEERING

ER&DCI INSTITUTE OF TECHNOLOGY

THIRUVANANTHAPURAM

2022

CERTIFICATE

This is to certify that the design report entitled **Data Anonymization** submitted by **Athul C K**, to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Cyber Forensics and Information Security is a bonafide record of the project work carried out by him under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Internal Supervisor(s)

Dr. Dittin Andrews

Ms. Dhanalakshmi M P

Harsha

External Supervisor

Ms. Harsha Gopalakrishnan

PG Coordinator

Dr. Alexander G

Principal

Ms. Rajasree S

ACKNOWLEDGEMENT

I take this opportunity to express my deep sense of gratitude and sincere thanks to all who helped me to complete the work successfully. My first and foremost thanks goes to God Almighty who showered in immense blessings on my effort.

I wish to express my sincere thanks to Ms. Rajasree S, Principal, ER&DCI Institute of Technology for providing me with all the necessary facilities and support.

I would like to express my sincere gratitude to Dr. Dittin Andrews, Ms. Harsha Gopalakrishnan and Ms. Dhanalakshmi M P, the internal project guide, for providing me with guidance and facilities for the mini-project. I also express my sincere gratitude to our project coordinator Ms. Gayathri S for her cooperation and guidance for preparing and presenting this report.

I wish to express my sincere gratitude towards all the teaching and non teaching staff members of our Department.

Finally I thank my parents, all my friends, near and dear ones who directly and indirectly contribute to the successful of this work.

ATHUL C K

ABSTRACT

In this digital age where data is valued at the highest, we need to ensure that the data collected by organizations does not contain personal data which may violate the privacy of the users. Many tech giants collect tremendous amounts of digital data from its users for business intelligence, analytics, training their recommendation systems and for much more. Researchers and the scientific community are also very much interested in these datasets since it can be used to solve many real-world problems, but the privacy regulations prevent the sharing of sensitive data to third-party organizations. Data anonymization is a way to preserve the utility of data and at the same time protect the privacy of users. It is defined as the process by which personal data is altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party. Privacy-preserving regulations like EU-GDPR, HIPAA, PCI-DSS among others, specify that datasets containing personally identifiable information (PII) must be properly anonymized before sharing it or making it available to the public. The General Data Protection Regulation (GDPR) permits companies to collect anonymized data without consent, use it for any purpose, and store it for an indefinite time as long as companies removes all the personal identifiers from the data. The lack of a proper definition of what a completely anonymized dataset is, and a robust algorithm which anonymizes the data without compromising the utility of data poses a problem to the scientific community. There has been multiple instances in which datasets were released to the public after performing what was believed to be proper data anonymization procedures but later found to be victims of de-anonymization. These cases directly points to the weaknesses in the anonymization process. In this study, industry-standard anonymization methods like k-anonymity, l-diversity, t-closeness, and differential privacy will be discussed, along with some new methods of anonymization and common industry practices.

Contents

List of Figures

List of Tables

ABBREVIATIONS

AOL	America On-Line
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
BK	Background Knowledge
CSV	Comma-Seperated Values
DISHA	Digital Information Security in Healthcare Act
DM	Discernibility Metric
DP	Differential Privacy
EU-GDPR	European Union-General Data Protection Regulation
GLP	Global Loss Penalty
HIPAA	Health Insurance Portability and Accountability Act
ISO/TS	International Organization for Standardization/Technical Specifications
ITM	Information Theocratic Metrics
JSON	JavaScript Object Notation
KL-divergence	Kullback–Leibler divergence
NLP	Natural Language Processing
NMI	Normalized Mutual Information
OS	Operating System
OSINT	Open Source Intelligence
PAN	Permanent Account Number
PCI-DSS	Payment Card Industry-Data Security Standards
PII	Personally Identifiable Information
PPDP	Privacy Preserving Data Publishing
QI	Quasi Identifier
SQL	Structured Query Language
TDA	Test-Driven Anonymization
UML	Unified Modelling Language
UUID	Universally Unique Identifier
XML	Extensible Markup Language

Chapter 1

INTRODUCTION

1.1 General Background

Humans have perfected the art and science of integration of data to our daily lives. We have access to data in amounts to a value that lies beyond the boundaries of human conception and at the same time we are well-occupied with the tools and techniques to manipulate the data in whichever way we like. Ubiquitous computing, higher bandwidth and extended connectivity has facilitated in a form of data collection and processing which is hitherto undreamt of. Data owners like banks, hospitals, insurance companies, social networks and other tech giants collect huge amounts of digital data from its users for a variety of purposes. This collected data often contains information about the users' activities, demographics, finance, hobbies, location, perceptions, interests, preferences, political and religious views, online communities, medical data and opinions. Organizations, enterprises and even the scientific community are very much interested in such datasets but are unable to lay their hands on these because privacy regulations prevent sharing of such sensitive data to third-party organizations. This forms the need for a solution which helps to share such datasets without compromising user's privacy. The solution must also be able to strike the perfect balance where the privacy of users are not compromised while preserving the utility of data. Data anonymization is a way in which this can be achieved. Privacy preserving data publishing (PPDP) provides set of models, tools, and methods to safeguard against the privacy threats that emerge from the data releasing with data miners or analysts. We require practical anonymization solutions to tackle this problem and also to protect data against the significant rise in privacy breaches across the globe. Data anonymization is the process by which personal data is altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party. It is aimed at protecting the privacy of users, preventing unintended disclosures, and preventing adversaries from reverse-engineering the data.

International data protection and privacy regulations like EU-GDPR, HIPAA, PCI-DSS among others, specify data anonymity as a requirement, which also heavily influenced Indian regulations like the Personal Data Protection Bill, 2019 [1] (aka Data Protection Bill, 2021) and the DISHA Act for health care data, just to mention a few. GDPR permits companies to collect anonymized data without consent, use it for any purpose, and store it for an indefinite time as long as companies remove all identifiers from the data. The lack of a proper definition of what a completely anonymized dataset is, and a robust algorithm which anonymizes the data without compromising the utility of data poses a problem to the technical community.

Google collects data from the Android OS, google analytics, and Chrome browser, Microsoft collects data from the Windows OS and applications on it, Uber collects data from its application, Amazon collects data from its site and application etc. All these data collected represents individuals across the globe and their preferences. These data points collected have a personal angle, and thus can be regarded as private information. This is done for a reason because companies need to know what a user wants, so that they can cater to the user's needs in a better way. Although the data is collected by these tech giants, it is not only these companies that wants this data. A lot of other parties are also interested in these datasets like advertisers, research scholars, developers, business consulting firms etc.

The ISO/TS 25237:2008 defines anonymization as the process that removes the association between the identifying data set and the data subject. The process of anonymization is sometimes also termed as de-identification, which means the process of removing, obscuring, redacting or delinking all personally identifiable information from an individual's digital health data in a manner that eliminates the risk of unintended disclosure of the identity of the owner and such that, if necessary, the data may be linked to the owner again. [2]

Anonymization is achieved by carefully threading the line between maintaining privacy and reducing information loss. The dataset must not loose its utility during the anonymization process, but also must loose all the PIIs and other information which could potentially lead to privacy violation. Also, the problem to re-identification still persists. An adversary must not be able to de-identify a properly anonymized dataset. Privacy preserving data mining and publishing methods have been attracting attention of the research community for some time, and many studies have been conducted on the problem. [3]

Personally Identifiable Information (PII) means any information that can be used to uniquely identify, contact or locate an individual, or can be used with other sources to uniquely identify a person. [2] A more detailed list of PIIs are provided in the appendix. De-anonymization is a reverse engineering process in which de-identified data are cross-referenced with other data sources to re-identify the personally identifiable information. This could occur if a de-identification process had not been not successfully performed, or had not been undertaken in the first place.

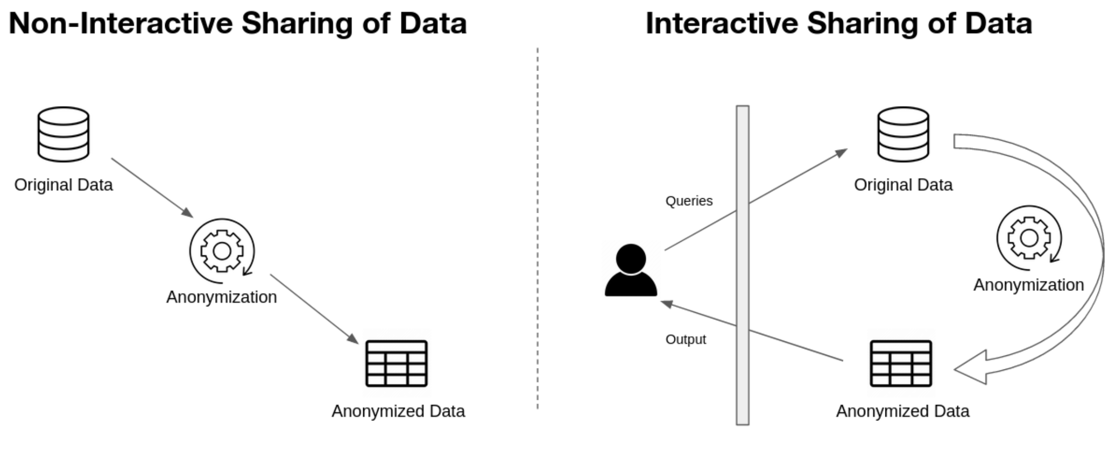


Figure 1.1: Two major data sharing paradigms

The process of sharing of sensitive data can be classified into two major methods. In non-interactive method, the data owner publishes the complete dataset in an anonymized form after applying some modifications on the original data. However, in the interactive setting, the data owner does not publish the whole data set in a sanitized form. Instead, data owner provides an interface to the data miners through which they may pose different statistical queries about the related data and get (possibly noisy) answers. The k-anonymity model and its ramifications are most widely used in the non-interactive setting of PPDP. These approaches apply some modifications on the original values of quasi identifiers, and protect the user's privacy by making information less- specific. The differential privacy (DP), and DP based approaches are mostly used in an interactive setting of PPDP.

Chapter 2

PROBLEM STATEMENT

Data anonymization as a topic is quite broad in the sense that it contains several problems statements in itself. Some of the major ones are described below:

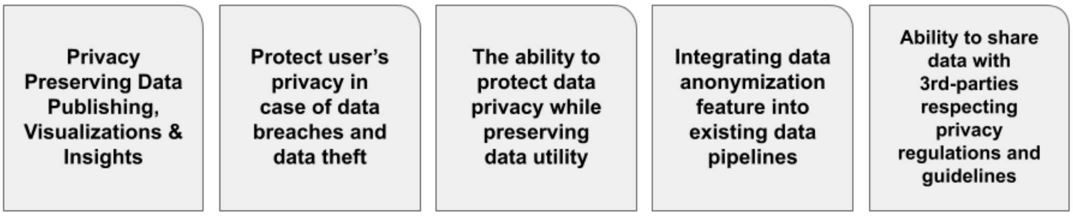
- **Privacy Preserving Data Publishing, Visualizations and Insights**: Privacy Preserving Data Publishing (PPDP) techniques focuses on various ways in which data can be published to the general public without violating the privacy of the users on the dataset. Not all applications needs the entire dataset to be published or shared with. Often, the requirement is just for some visualizations or insights made from the underlying dataset to understand the general trend. These visualizations and insights must also not violate privacy of the users.

- **Protecting privacy in case of data breaches and thefts**: We have been hearing news of data breaches and data theft from time to time. The leaked data could potentially harm the users' privacy if the attackers are capable to map the users' identity to the sensitive data property values. Storing and working with properly anonymized data removes the issue of privacy breach in the unfortunate event of a data leak.

- **Ability to share data with 3rd-parties respecting privacy regulations and guidelines**: Privacy regulations like EU-GDPR, Personal Data Protection Bill (2019), DISHA Act (2018), HIPAA, PCI-DSS and others aims to protect the identity of the user and uphold the privacy of the individuals. Many of these regulations require companies to anonymize their data so that even in case of a data leak, the private information of the users will not be exposed. Also, there are guidelines preventing the datasets to be shared with third-parties without proper consent form the users, and also the dataset which is being shared must not reveal the identity of the users in the dataset. Many of the advertising and business analytics companies and also the research community are very much interested in these datasets, but these privacy regulations prevents sharing unanonymized data.

- **Maintaining the balance between data privacy and utility**: There arise a need for a proper data anonymization algorithm which can adhere to the privacy regulations by protecting the privacy of the users without actually degrading the utility of the dataset. As we turn the knob to increase data privacy, this would in turn decrease the utility of the data available. Similarly, too much anonymization may end up generating a very generalized and random dataset which is of no use to the consumer.

- **Integrating data anonymization into existing pipelines**: As the legal requirements develops, there will be a rise in need of flawless integration of data anonymization to existing data pipelines available. Tools, APIs or even anonymization as a service can be considered an effective solution to the problem.



Privacy Preserving Data Publishing, Visualizations & Insights

Protect user's privacy in case of data breaches and data theft

The ability to protect data privacy while preserving data utility

Integrating data anonymization feature into existing data pipelines

Ability to share data with 3rd-parties respecting privacy regulations and guidelines

Figure 2.1: Various problem statements in data anonymization

2.1 Scope

Data may be available in many formats. Structured data may be of the form of CSVs, relational and no-SQL databases, graph data, JSON, XML etc. Unstructured data will be in a textual fashion. There is also multimedia data like photos, videos, audio data etc. In a way, all categories of data needs to anonymized for ideal security measures. But, as far as this work is considered, the focus is on relational data anonymization.

Chapter 3

LITERATURE SURVEY

3.1 Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey

The work done by Abdul Majeed and Sungchang Lee [3] has served to be the baseline of this study. Many concepts and ideas were introduced to this from the aforementioned study. Although their study focuses on both relational and structural datasets, the scope of this project was limited to relational data structures only. The extensive literature survey conducted by the authors referenced over 300 papers and churned out the essence into their work. They covered statistical privacy models like k-anonymity, l-diversity and t-closeness along with differential privacy model.

3.2 Analysis of Data Anonymization Techniques

This work [4] focuses on the GDPR requirement of data anonymization and how it differs from pseudo-anonymization. The work studies the naive anonymization operations like suppression, substitution, shuffling(permutation), noise addition, and generalization while specifying the risk associated with each. It also looks into various anonymization tools like ARX, µ-Argus, SDCMicro and Privacy Analytics Eclipse.

3.3 Test-Driven Anonymization in Health Data: A Case Study on Assistive Reproduction

The paper [5] evaluates their previous approach Test-Driven Anonymization (TDA) with a real-world health dataset from the Spanish Institute for the Study of the Biology of Human Reproduction (INEBIR). TDA aims to anonymize the data incrementally by testing each of these anonymization efforts to obtain the dataset that achieves a trade-off between the functional (functional suitability) and non-functional (privacy) quality. Their study revealed that for this specific dataset, k-anonymity with a k value of 16 produces best results which were able to preserve the anonymity as well as not to loose too much utility to cause error in the AI model during the training phase.

3.4 Privacy preserving data publishing and data anonymization approaches: A review

In the survey review paper [6], they surveyed various developments and algorithms suggested in the field of PPDP. Although the main objective is the transformation of original data into a form from which deduction of the record owners' sensitive information can be prevented, they presented the advantages and disadvantages of 11 various algorithms suggested by various authors in last decade. Most of the suggested algorithms assumed a single data release from a publisher, thus only protected the data up to the first release or the first recipient.

3.5 Simple Demographics Often Identify People Uniquely

The work by L. Sweeny [7] is regarded as one of the core baseline and a fundamental concept on the field of data anonymization. The work gave proof to the fact that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on 5-digit ZIP, gender, date of birth.

3.6 Who's Watching? De-anonymization of Netflix Reviews using Amazon Reviews

This work [8] describes the de-anonymization of the famous Netflix Prize dataset by MIT researchers, which was then believed to be an anonymized dataset published by Netflix. Even though all personal identifiable information was removed and the data was slightly perturbed, the authors were able to perform a membership attack on the Netflix Prize dataset by correlating it with the publicly available Amazon reviews.

3.7 Ministry of Health & Family Welfare Government of India. "Digital Information Security in Healthcare, Act (DISHA)", 2017.

This [2] is an Act to provide for establishment of National and State eHealth Authorities and Health Information Exchanges to standardize and regulate the processes related to collection, storing, transmission and use of digital health data; and to ensure reliability, data privacy, confidentiality and security of digital health data and such other matters related and incidental thereto. It specifies certain requirements regarding data anonymization along with a non-exclusive list of PIIs.

3.8 The EU Working Party on the Protection of Individuals with Regard to the Processing of Personal Data., Opinion 05/2014 on Anonymisation Techniques, 2014.

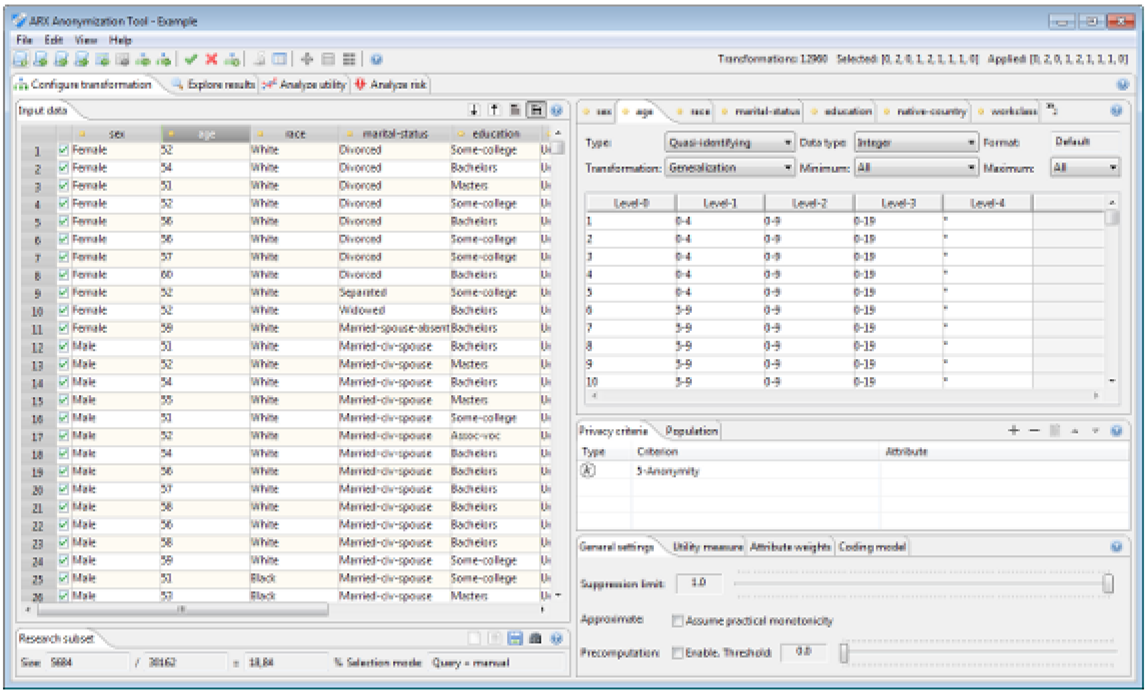
This document [9] describes in depth on many anonymization techniques including randomization, noise addition, permutation, differential privacy, generalization, k-anonymity, l-diversity, t-closeness and pseudonymization, and also discusses the guarantees, pitfalls and shortcoming of each.

Chapter 4

EXISTING SYSTEM

4.1 ARX Data Anonymization Tool

ARX is a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of privacy and risk models, various methods for transforming data and for analyzing the usefulness of output data. The software has been used in a variety of contexts, including commercial big data analytics platforms, research projects, clinical trial data sharing and for training purposes. ARX is able to handle large datasets on commodity hardware and it features an intuitive cross-platform graphical user interface.



Transformations: 12960 Selected: [0, 2, 0, 1, 2, 1, 1, 1, 0] Applied: [0, 2, 0, 1, 2, 1, 1, 1, 0]

		sex	age	race	marital-status	education	
1	☑	Female	52	White	Divorced	Some-college	U
2	☑	Female	54	White	Divorced	Bachelors	U
3	☑	Female	51	White	Divorced	Masters	U
4	☑	Female	52	White	Divorced	Some-college	U
5	☑	Female	56	White	Divorced	Bachelors	U
6	☑	Female	56	White	Divorced	Some-college	U
7	☑	Female	57	White	Divorced	Some-college	U
8	☑	Female	60	White	Divorced	Bachelors	U
9	☑	Female	52	White	Separated	Some-college	U
10	☑	Female	52	White	Widowed	Bachelors	U
11	☑	Female	59	White	Married-spouse-absent	Bachelors	U
12	☑	Male	51	White	Married-civ-spouse	Bachelors	U
13	☑	Male	52	White	Married-civ-spouse	Masters	U
14	☑	Male	54	White	Married-civ-spouse	Bachelors	U
15	☑	Male	55	White	Married-civ-spouse	Masters	U
16	☑	Male	51	White	Married-civ-spouse	Some-college	U
17	☑	Male	52	White	Married-civ-spouse	Assoc-voc	U
18	☑	Male	54	White	Married-civ-spouse	Bachelors	U
19	☑	Male	56	White	Married-civ-spouse	Bachelors	U
20	☑	Male	57	White	Married-civ-spouse	Bachelors	U
21	☑	Male	58	White	Married-civ-spouse	Bachelors	U
22	☑	Male	56	White	Married-civ-spouse	Bachelors	U
23	☑	Male	58	White	Married-civ-spouse	Bachelors	U
24	☑	Male	59	White	Married-civ-spouse	Some-college	U
25	☑	Male	51	Black	Married-civ-spouse	Some-college	U
26	☑	Male	53	Black	Married-civ-spouse	Masters	U

Research subset

Size: 5684 / 30162 = 18,84 % Selection mode: Query = manual

Type: Quasi-identifying Data type: Integer Format: Default

Transformation: Generalization Minimum: All Maximum: All

Level-0	Level-1	Level-2	Level-3	Level-4
1	0-4	0-9	0-19	*
2	0-4	0-9	0-19	*
3	0-4	0-9	0-19	*
4	0-4	0-9	0-19	*
5	0-4	0-9	0-19	*
6	5-9	0-9	0-19	*
7	5-9	0-9	0-19	*
8	5-9	0-9	0-19	*
9	5-9	0-9	0-19	*
10	5-9	0-9	0-19	*

Type	Criterion	Attribute
(k)	5-Anonymity	

Suppression limit: 1.0

Approximate: □ Assume practical monotonicity

Precomputation: □ Enable. Threshold: 0.0

Figure 4.1: ARX Anonymization Tool

ARX supports a variety of privacy models like k-anonymity, k-map, average risk, population uniqueness, sample uniqueness, l-diversity, t-closeness, δ-disclosure privacy, β-likeness, δ-presence, profitability and differential privacy.

4.2 Microsoft Presidio

Presidio helps to ensure that sensitive data is properly managed and governed. It provides fast identification and anonymization modules for private entities in text and images such as credit card numbers, names, locations, social security numbers, bitcoin wallets, US phone numbers, financial data and more.

It allows organizations to preserve privacy in a simpler way by democratizing de-identification technologies and introducing transparency in decisions. Supports extensibility and customizability to a specific business need and facilitates both fully automated and semi-automated PII de-identification flows on multiple platforms.

4.3 Amnesia

The Amnesia anonymization tool is software written in Java and JavaScript and should be used locally for anonymizing personal and sensitive data. The basic idea behind anonymization is that users load a file containing personal data (original data) to Amnesia, and Amnesia transforms it into an anonymous dataset, which can then be stored locally. The transformation is guided by user selections and provides an anonymization guarantee for the resulting dataset. Amnesia currently supports k-anonymity and km-anonymity guarantees. Performs research and share your results that satisfy GDPR guidelines by using data anonymization algorithms. The tool supports pseudo-anonymization, masking, K-anonymity, Km-anonymity and demographic statistics.

4.4 μ-ARGUS & τ-ARGUS

μ-ARGUS is a software program designed to create safe micro-data files. The CASC-project took a previous version of μ-ARGUS, developed a.o. during the SDC-project and prototypes before that, as a starting point. At the end of the CASC-project version 4.0 was available. The ESSnet-project has made it possible to further extend μ-ARGUS. This resulted in Version 4.2. Partly funded by Eurostat, a project was started on 20 December 2012 to port the at that time most recent version of μ-ARGUS to an Open Source version.

The resulting version should contain the possibility to be run on a Windows platform as well as on a Linux/Unix platform. As of version 5.1.6 we will only provide 64 bit versions with bundled JRE (Zulu Open JRE).

4.5 Diffix

Diffix is an algorithm for anonymizing structured data. It was jointly developed by Aircloak GmbH and the Max Planck Institute for Software Systems. Diffix combines the three most common anonymization mechanisms, generalization, noise, and low-count suppression. It automatically applies these mechanisms as needed on a query-by-query basis to minimize noise while ensuring strong anonymity.

4.6 Anonimatron

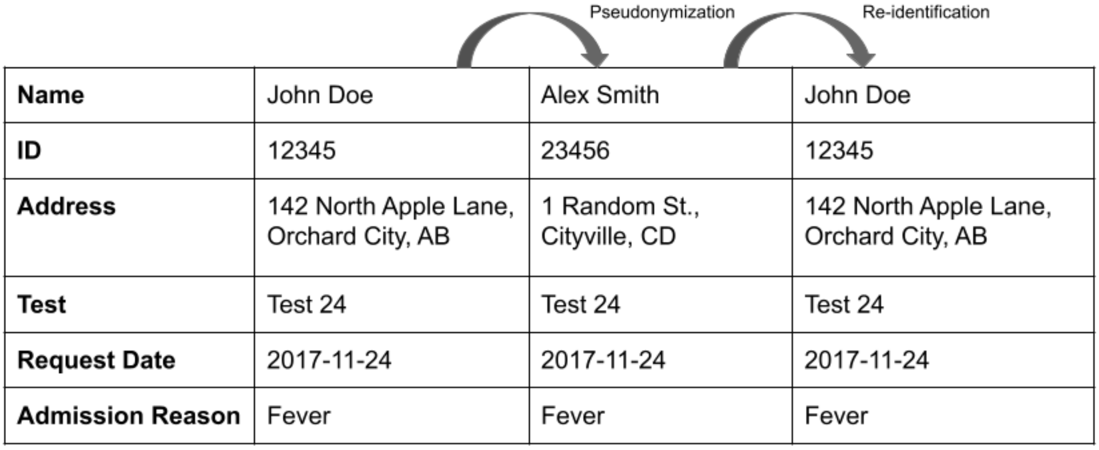
Anonimatron is a free, extendable, open source data anonymization tool for anonymizing databases and files. It supports Oracle, PostgreSQL and MySQL out of the box. Users can add their own anonymization handlers and can generate fake email addresses, fake Roman names, and UUID's out of the box. It is easy to configure, runs on Windows, Mac OSX, Linux derivatives and is free of charge.

Chapter 5

SYSTEM STUDY

5.1 Pseudonymization

Pseudonymisation consists of replacing direct identifiers in a record by another. The natural person is therefore still likely to be identified indirectly; accordingly, pseudonymization when used alone will not result in an anonymous dataset. Pseudonymisation is not considered to be a method of anonymization as it merely reduces the linkability of a dataset with the original identity of a data subject, and is merely suggested as a useful security measure.The result of pseudonymisation can be independent of the initial value (as is the case of a random number generated by the controller or a surname chosen by the data subject) or it can be derived from the original values of an attribute or set of attributes e.g. a hash function or encryption scheme. The example shown in Fig.5.1 depicts a simple and naive pseudonymization process.


Pseudonymization
Re-identification

Name	John Doe	Alex Smith	John Doe
ID	12345	23456	12345
Address	142 North Apple Lane, Orchard City, AB	1 Random St., Cityville, CD	142 North Apple Lane, Orchard City, AB
Test	Test 24	Test 24	Test 24
Request Date	2017-11-24	2017-11-24	2017-11-24
Admission Reason	Fever	Fever	Fever

Figure 5.1: Pseudonymization Example

A specific pitfall is to consider pseudonymised data to be equivalent to anonymised data because they continue to allow an individual data subject to be singled out and linkable across different data sets. Pseudonymity is likely to allow for identifiability, and therefore

stays inside the scope of the legal regime of data protection. This is especially relevant in the context of scientific, statistical or historical research.
A typical instance of the misconceptions surrounding pseudonymisation is provided by the well-known "AOL (America On Line) incident". In 2006, a database containing twenty million search keywords for over 650,000 users over a 3-month period was publically released, with the only privacy preserving measure consisting in replacing AOL user ID by a numerical attribute. This led to the public identification and location of some of them. Pseudonymised search engine query strings, especially if coupled with other attributes, such as IP addresses or other client configuration parameters, possess a very high power of identification. The most commonly found naive methods of pseudo-anonymization are:

- Encryption
- Hash Function
- Tokenization

5.2 Relational Data Anonymization

Any data that can be represented in tabular format falls under relational data classification. That is, there is a relation between the fields in various columns across a data row. Since most the databases already uses relational data model, anonymization of such datasets naturally falls under the requirement of the industry.

When it comes to relational data anonymization, it is not as simple as just removing the PII attributes and publishing the dataset. There are multiple incidences in which simple removal of what was considered to be PIIs ended up badly and adversaries figuring out the private data and finally to revealing the identity of the users [8]. These incidents wrere in direct correlation of what was published in this paper [7] which proves that about 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on their 5-digit ZIP, gender, date of birth. So, a detailed analysis of the attributes needs to be done in case of relational data anonymization.

One can classify the attributes in a relational data as the following:

- **Direct identifiers** serve to "uniquely" identify an individual, and include data elements such as Social Security Numbers, Tax ID numbers, Passport numbers, full names or addresses. These are mainly refered to as PIIs.
- **Indirect or quasi identifiers**, while not unique to an individual, can be combined with other indirect identifiers to identify an individual among a set of individuals. Indirect identifiers include items such as zip code, birth date, IP address, etc.
- **Sensitive data identifiers** are values which the user don't want to be publicly disclosed, but many-a-times these may be the attributes under study. So, removing or manipulating such attributes will significantly impact the utility of the dataset. Examples include salary, disease info, political/religious preferences etc.
- **Other personal data** elements may be associated with multiple individuals, such as level of education, area of study, or communication preferences, and within a single data set a combination of such elements often does not allow the identification of a single individual.

When data have been appropriately manipulated, combined or aggregated (perhaps in census data or survey results) they typically can no longer be linked to any individual, and are considered anonymized. Some data elements (such as weather) are simply not related to individuals, and would be considered as non-personal information. The table 5.1 describes how various user attributes needs to be handled in an anonymization process.

Data sets containing either direct or indirect identifiers are generally perceived to be more useful for research or analytics, and typically present greater risks to individual privacy. Historically, in order to reduce such privacy risks, the following techniques have been used:

1. Deletion, redaction or obfuscation: Direct identifiers are covered, eliminated, removed or hidden. These techniques are difficult to accomplish well, particularly on unstructured data, and use of unsophisticated techniques may enable easy re-identification. Example: Jane Doe – DOB 8/15/1970 – St. Louis → Jane Doe – DOB 8/15/1970 – St. Louis

2. Pseudonymization: Information from which direct identifiers have been eliminated,

Type	**Description**	**Examples**	**Action**
Direct Identifiers	Can uniquely and directly identify an individual/user	Name, Social Security Number, Email, Phone Number	Removed
Quasi Identifiers	Can be linked with auxiliary information to reveal someone's identity	Age, Gender, Race, Zip Code	Generalized
Sensitive Attributes	Values which the user don't want to be publicly disclosed	Salary, Disease info, Political/Religious preferences	Retained
Non-Sensitive Attributes	All attributes other than Direct, Quasi and Sensitive attributes	Height, weight, eye color etc.	Usually not collected

Table 5.1: Description about the types of user's attributes and their handling in an anonymization process.

transformed or replaced by pseudonyms, but indirect identifiers remain intact. Re-identification may occur where there is failure to secure the pseudonymization method or key used, and/or when reverse engineering is successful. Example: Jane Doe – DOB 8/15/1970 – St. Louis → ID:TRXD 8/15/1970 St. Louis

3. De-identification: Direct and known indirect identifiers (perhaps contextually identified by a particular law or regulation, i.e. HIPAA) have been removed or mathematically manipulated to break the linkage to identities. Example: Jane Doe – DOB 8/15/1970 – St. Louis → Female 1970 Missouri

4. Anonymization: Direct and indirect identifies are removed or manipulated together with mathematical and technical guarantees, often through aggregation, in order to prevent re-identification. Anonymization is intended to be irreversible. Example: Jane Doe

– DOB 8/15/1970 – St. Louis → Female Adult Missouri

Note that encryption is sometimes inaccurately thought of as an obfuscation or de-identification technique. However, it is not a such a technique, but rather is a security measure intended to protect the personal data that may contain any combination of identifiable data elements. Apart from all the various methods described to anonymize the relational datasets, what is truly the need is to form a mathematical model which can provide a measure for the anonymity parameter. This is where syntactic anonymization models along with differential privacy comes into play.

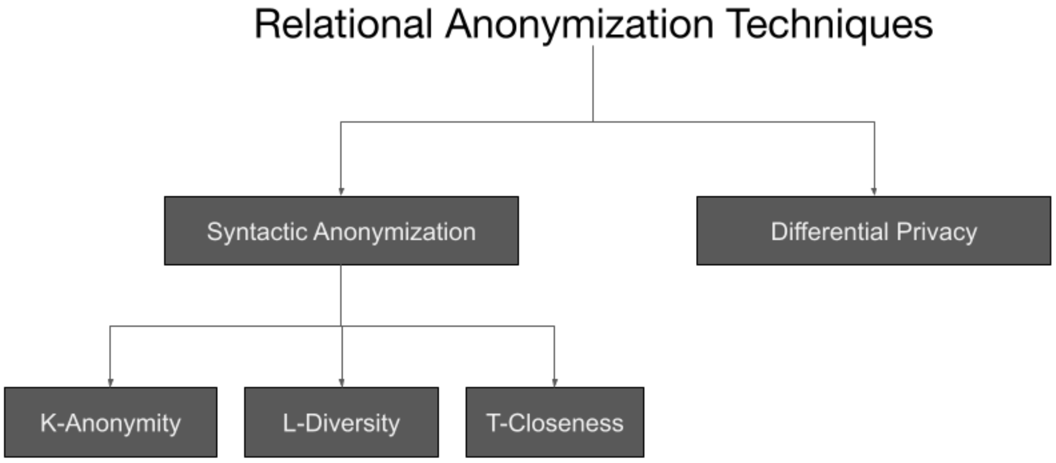


Figure 5.2: Common relational anonymization techniques

5.2.1 k-Anonymity

In the k-anonymity privacy model, the k is a number that represents the size of a group of users in the dataset who belongs to a single equivalent class. If for any individual in the data set, there are at least k-1 individuals who have the same properties, then we have achieved k-anonymity for the data set. The k-anonymity model protects user's privacy by placing at least k users in an equivalence class with same QI values. Hence, the probability of re-identifying someone from the anonymized data becomes $1/k$.

5.2.2 l-diversity

l-diversity anonymity guarantees l different values for each group's sensitive attributes. Thus, an attack can recognize a user's sensitive information with maximum probability of $1/l$. Well representation of sensitive data in every group is the main objective of l-diversity. According to this privacy model, a class satisfies l-diversity property if there are at least l "well-represented" values for the sensitive attribute.

5.2.3 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.A table T satisfies t-closeness if its records/tuples are split into equivalent classes such that the distribution of sensitive attributes in the whole table and the equivalent classes of a t-close table are within t distance units of each other. The value of t can be determined by considering the protection level. The t-closeness privacy model significantly improves the user's privacy, but it severely reduces the utility of the released data.

5.2.4 Differential Privacy

It has been reported in the literature that DP provides a mathematically provable guarantee on privacy preservation against many privacy attacks such as differencing, linkage, and reconstruction attacks. Before defining differential privacy model, we need to understand what neighbouring datasets are. Neighbouring datasets (say, D1 and D2) differ in one and only one record. Thus differential privacy can be defined as: given a dataset D1, and a neighbouring dataset D2, we define a random function F with a range S, defined as S $\subseteq$ Range(F), satisfies DP if:

$$\texttt{Pr[F(x)} \in \texttt{S]} \leq \texttt{e}\epsilon\ \texttt{Pr[F(y)} \in \texttt{S] +} \delta$$

variable x, y $\in$ D1, D2;
ϵ is a parameter;

δ indicates a degree of the relaxation.

DP mathematically ensures that any analyst/data-miner seeing the result of a differentially private analysis will make the same conclusion about any individual's private information, whether or not that individual's private information is included in the input to the analysis.

5.3 Evaluation Metrics

Evaluation metrics are used to evaluate various aspects of the anonymized data. Usually, the utility and risk associated with the anonymized data are examined to make sure that the output dataset serves its expected purpose well while making sure that the risk of de-identification of the data subjects remain minimum. The following subsections details the two metrics.

5.3.1 Privacy Evaluation Metrics

There exist multiple ways to quantify the privacy protection offered by an anonymization algorithm. The five common methods that are employed to evaluate the effectiveness of any anonymization algorithm in terms of privacy protection are detailed below.

1. Calculating the probability of successful matches based on the QI values from both original & anonymized dataset

Anonymous and original dataset linking and calculating the probability of successful matches based on the QI values in both these datasets. The probability value tells the amount of privacy protection an anonymization algorithm will offer during published data analysis. In this case, it is assumed that attackers may have access to the excessive amount of auxiliary information from some external sources (i.e., voter list, online repositories, e-commerce sites etc.) to launch identity and attribute disclosure attacks.

2. Evaluation in presence of some background knowledge

In this privacy evaluation metric, the data owner assumes that attacker may possess the true information (i.e., age and gender of a Bob) about some users and he/she can explore only the particular equivalent classes to infer private information of relevant user/users. To evaluate effectiveness of algorithm in this regard, the data owner can pick some instances from the original data and can evaluate the anonymization algorithm privacy protection level by matching.

3. Evaluation with the help of privacy-sensitive (PS) rules
In this case, the data owner can construct certain rules to evaluate privacy protection. For example, how many people having age greater than 40 suffer from this disease(i.e., cancer). The sensitive knowledge pattern revelation, and attribute and identity disclosure of multiple users through PS rules have a wide range of negative consequences on people's life.

4. Prediction about the users Sensitive Attributes through the existence of private and public profiles in Social Networks
In this case, the data owner can quantify the amount of protection level of an algorithm assuming that either partial or full user's original data is known to the attacker as some users willingly publish their QIs over different Social Networks. The factual information that can be learned through the knowledge gained from the data to invade unknown users' privacy can be used to evaluate an anonymization effectiveness.

5. Privacy protection evaluation in the presence of malicious users' in a dataset
In this case, the data owner can classify some of the tuples as malicious and can calculate the similarity with other (i.e., non-malicious) users to quantify the privacy protection level. In some cases, the sensitive queries and corresponding private information budget is also used for the evaluation of anonymization algorithms/models.

5.3.2 Utility Evaluation Metrics

During tabular data anonymization, the original QI's values are modified to fulfill the privacy needs, hence, the data utility degrades. There exist multiple ways to quantify the anonymous data utility offered by an anonymization algorithm. One can classify the metrics used for measuring anonymous data utility into the following two categories:

1. Special Purpose Metrics

The special purpose metrics use machine learning methods to measure the anonymous data quality. The most widely used special purpose metrics are, accuracy or error rate, F-measures, precision, and recall.

2. General Purpose Metrics

The general purpose metrics measure the information loss caused by modifying the original data. The most popular general purpose utility evaluation methods are, weighted certainty penalty, generalized information loss (GenILoss), discernability metric, minimal distortions, average equivalence class size ($\mathrm{C}_{\mathrm{AVG}}$), KL-divergence, granularity, query accuracy, global loss penalty (GLP), normalized mutual information (NMI), relative error (RE), and information theocratic metrics (ITM).

5.3.3 Risk Evaluation

There is a necessity to acknowledge the risks and threats associated with data anonymization as a process. There exists three classes of privacy threats that can occur during published data analysis, which are explained below:

• *Identity disclosure* (i.e., unique identification): It is a well-known privacy threat in the PPDP. It occurs when an adversary can correctly associate an individual in a privacy preserved published dataset. Generally, an attacker use the information gathered from external sources (i.e., voter registration list, online repositories, and factual information) to identify an individual uniquely.

• *Attribute disclosure* (i.e., private information disclosure): This type of privacy threat occurs when an individual is linked with the information about his/her SA. For example, the information can be the person's value for the sensitive attributes. This type of threat can be easily launched in imbalanced datasets (i.e., the datasets lacking heterogeneity in SA's values.).

- *Membership disclosure* (i.e., presence/absence disclosure): This threat occurs when an adversary can deduce that an individual's record is present/absent in the published dataset with a very high probability. Researchers have reported many interesting scenarios in which the protection from the membership disclosure is imperative.

For each dataset and for each anonymization model, these threats and how it affects the anonymized datasets needs to be evaluated and calculated before proceding with the public publishing of the datasets. Also, the below table 5.2 depicts various risks associated with the anonymization and pseudonymization methods considered.

	Singling out	Linkability	Inference
Pseudonymisation	Yes	Yes	Yes
Noise addition	Yes	May not	May not
Substitution	Yes	Yes	May not
Tokenization	Yes	Yes	May not
K-anonymity	No	Yes	Yes
L-diversity	No	Yes	May not
Differential Privacy	May not	May not	May not

Table 5.2: Strengths and Weaknesses of the Techniques Considered

As you can see in the table 5.2, there are risks associated with each and every algorithms discussed. There is no one-stop solution when it comes to anonymization, but still many seems to prefer differential privacy because of the mathematical security it provides. Tech giants like Google, Uber, Microsoft etc. has already adopted various implementations of differential privacy on their products.

Chapter 6

PROPOSED SYSTEM

The proposed system, as shown in Fig.6.1, undergoes a 3-stage process to properly anonymize the input data. The tool expects to receive the input data in a relational format, CSV for example. The input data then undergoes a series of data preprocessing steps to make sure that the data is fit to be fed into the anonymization process. Details of the preprocessing steps can be found under the methodology sub-section.

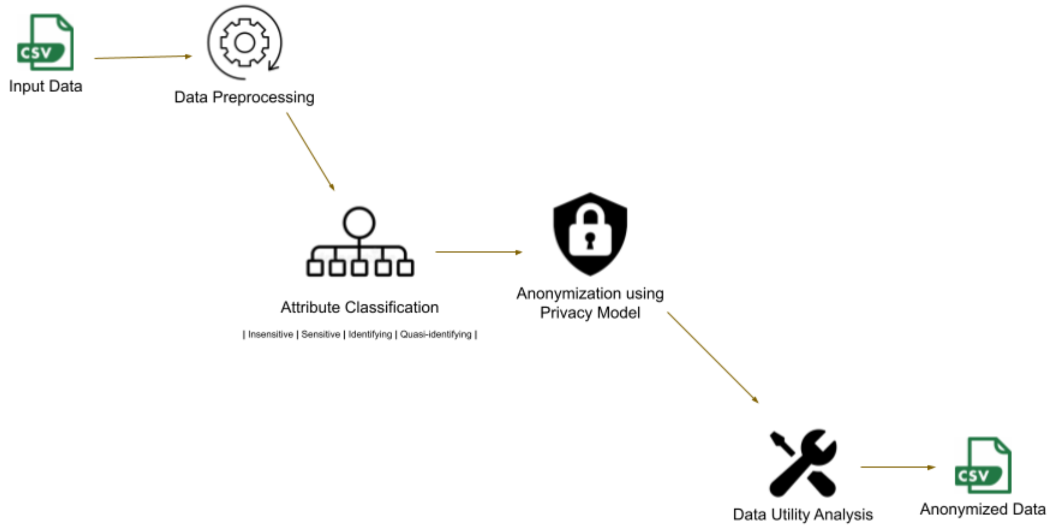


Figure 6.1: Block diagram of proposed system

The anonymization stage starts with an attribute classification process. Once the attribute are categorized according to its sensitivity, the user is prompted to select a required privacy model for anonymization. After the anonymization process, the utility of anonymized data needs to be evaluated to ensure that the data is still useful for the intended purpose. The last stage starts with the analysis of the data re-identification risk. We need to try various attacks an adversary can do to the resulting anonymized data to de-anonymize it. Once those risks, if any, are handled properly, the data can be exported to a data format the user needs.

6.1 Methodology

DATA PRE-PROCESSING

1. Outlier removal: Removal of outliers from the input dataset to make sure that the anonymization process can be performed effectively.

2. Datatype selection: Each attribute is given a certain datatype.

3. Bias removal: Some datasets may be biased from the initial stage. This is of importance because the anonymization process may not be that effective due to the inherent bias on the dataset.

ATTRIBUTE CLASSIFICATION

1. Sensitive: Sensitive attributes are values which the users don't want to be publicly disclosed. Exmples include salary details, disease information, political/religious preferences etc. These value are retained because the utility of the dataset may depend upon studying these attributes.

2. Insensitive: All attributes other than direct identifiers, quasi and sensitive attributes. Examples include height, weight, eye color etc. These values are not collected in most cases, and even if collected, they can be retained or removed because they may not influence the utility of the dataset or privacy of the users.

3. Identifiers (PIIs): Direct identifiers can uniquely and directly identify an individual/user. Examples include name, Social Security Number, email, phone numbers etc. The initial values are removed.

4. Quasi-identifiers: Such attributes an be linked with auxiliary information to reveal someone's identity. Examples include age, gender, race, zip code etc. Such values are generalized.

PRIVACY MODELS

1. k-anonymity

2. l-diversity

3. t-closeness

4. Differential Privacy

Refer Section 3.2 for more details.

UTILITY ANALYSIS

1. Information loss: We need to implement proper methods to measure the overall information loss incurred because of the anonymization process, so that we can minimize it.

2. Statistical measures: It needs to be noted that the anonymization process does not change the overall statistical properties of the dataset too much.

DE-IDENTIFICATION RISK ANALYSIS

Records with the highest de-identification risk needs to be subjected to further anonymization process to hold up the privacy promise. As of now, the risk of re-identification needs to be computed manually by the analyst.

Chapter 7

SYSTEM DESIGN

7.1 High-Level Design

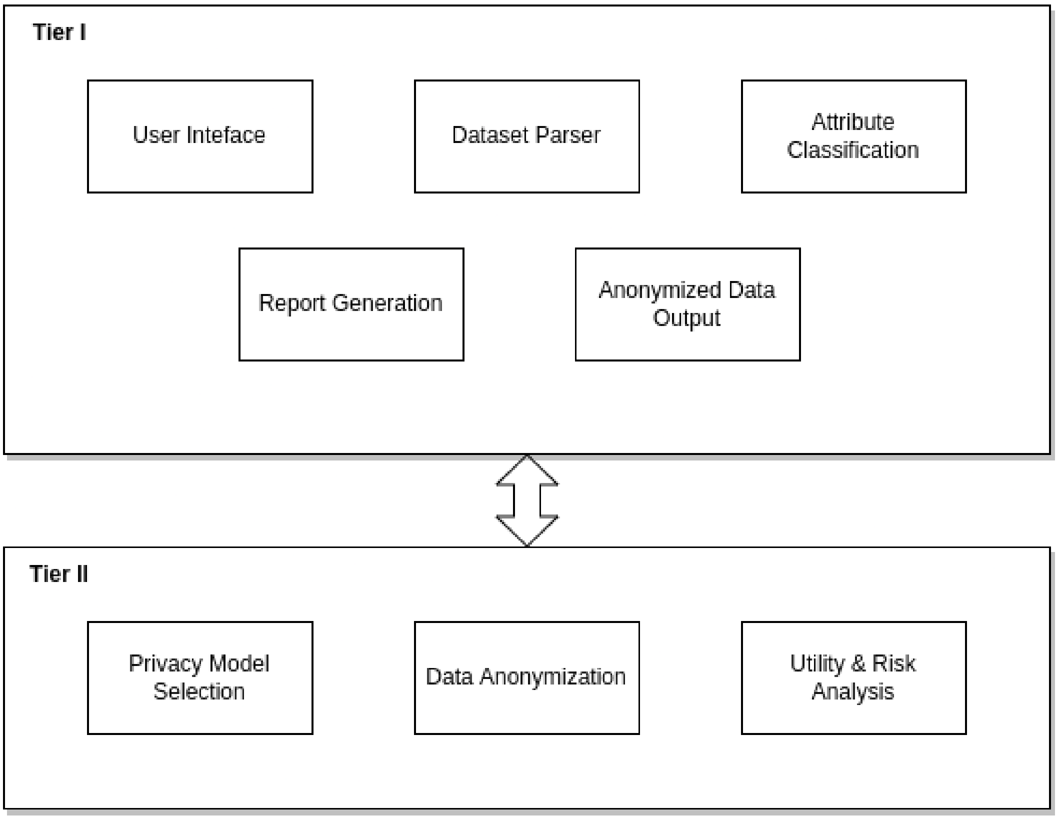


Figure 7.1: System Architecture

The architecture of the proposed application, shown in Fig.7.1, can be divided into two tiers. Tier I consists of the general user interface module, along with some other modules which directly interacts with the user and the data. The dataset parser module takes in the input data file and loads it into memory. It does data type detection. The attribute classification module aids the user to pick and choose which all attributes are direct identifiers, quasi-identifiers, sensitive and insensitive attributes. The anonymized data output module is used to export the anonymized data into a file. The report generation module generates an ASCII report specifying all the configuration parameters and anonymization details.

Tier II consists of 3 major modules. The privacy model selection module helps in selecting and configuring the privacy model according to the user's preference. This work mainly aims in implementing differential privacy model. The data anonymization module is where the core anonymization process takes place. The utility & risk analysis module helps the user to pick and choose required result with minimal risk and maximum utility.

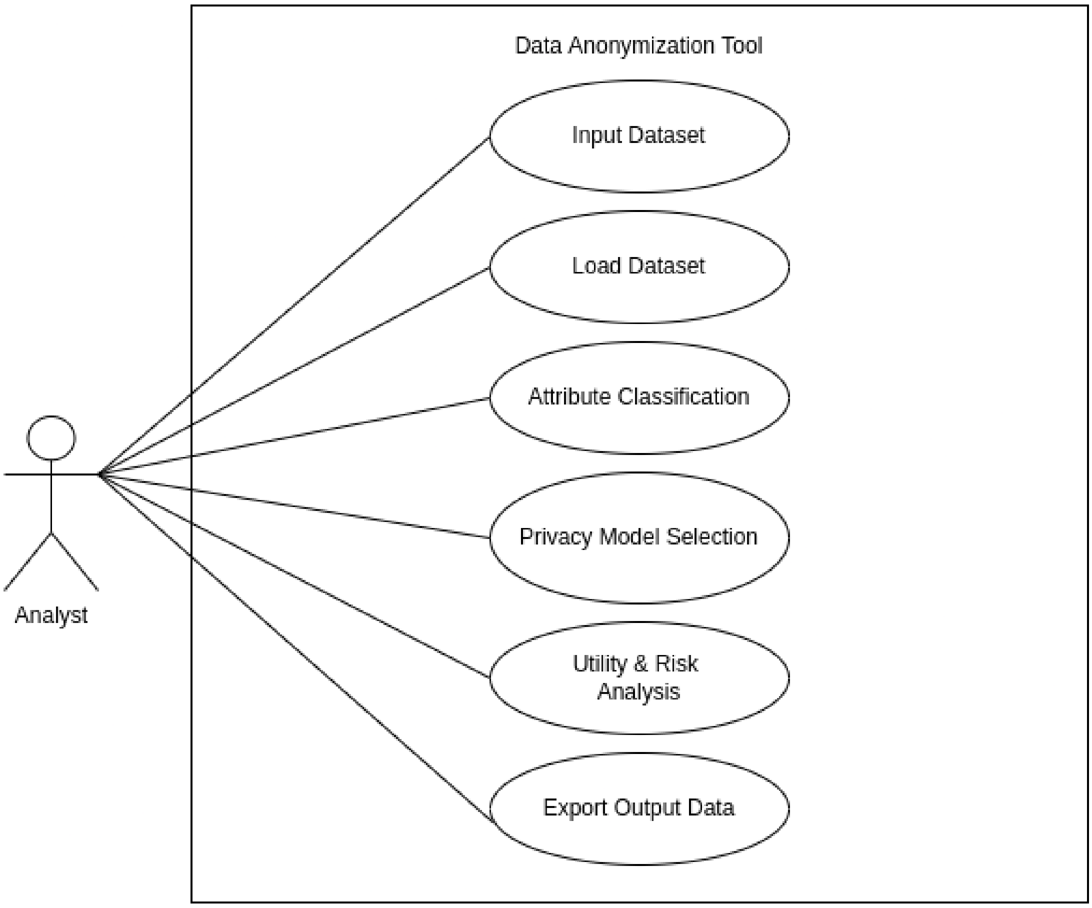


Figure 7.2: UML diagram

The Fig.7.2 shows the use case diagram of the application. The subject is the analyst who performs the anonymization process. The diagram illustrates various actions the analyst can perform on the application.

7.2 Low-Level Design

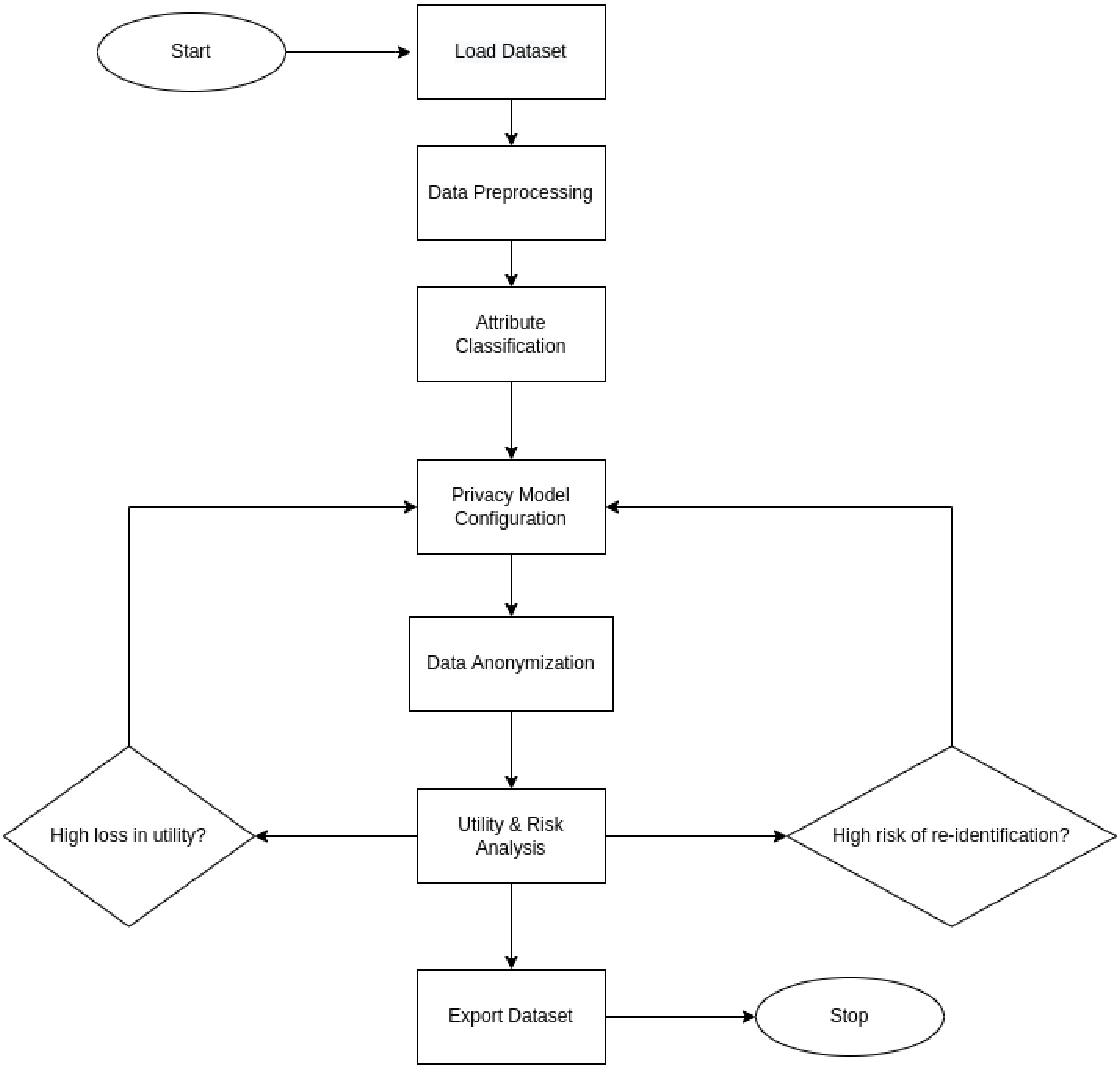


Figure 7.3: Flow diagram

The flow diagram shown in Fig.7.3 shows the overall flow of the anonymization process using the application. Once the anonymization process is completed, the user needs to evaluate the overall risk and utility of the output dataset and reconfigure the privacy model, if necessary. The application provides the user various options to tweak the privacy model to obtain a better output dataset which may have better utility and low risk of de-identification.

Chapter 8

HARDWARE & SOFTWARE REQUIREMENTS

The tool is written using Python programming language, and can be build for both Windows and Linux operating systems.

8.1 Hardware Requirements

CPU: Intel i3 or above.

Memory: 2GB or more.

Hard Disk: 50GB or more.

8.2 Software Requirements

Windows XP or above/Ubuntu (or any linux distro)

Python 3

Chapter 9

IMPLEMENTATION

9.1 Data Preprocessing

The first step involved after properly reading the dataset is data pre-processing. Datasets may contain NaN values, some fields may be missing values or some fields may be completely be absent. In such cases, the entire row is removed from the dataset. Datatype selection is also included in the pre-processing step, wherein each attribute is assigned a specific datatype. These can be strings, numbers, floating point numbers etc.

9.2 Attribute Anonymization

Given the dataset, we need to classify the attributes into one of the following categories and apply the anonymization procedure appropriately.

9.2.1 Direct-identifiers

Direct-identifiers are attributes which are very powerful in re-identifying a particular individual. Attributes like PAN number, Social Security Number, Phone number, email addresses etc. are very useful in pin-pointing a certain individual from a dataset with million records. Hence, leaving it in the anonymized dataset will be like inviting disaster. Hence, we are suppressing the direct-identifiers with the symbol '*'.

9.2.2 Quasi-identifiers & Sensitive attributes

Quasi-identifiers are anonymized using k-anonymity model whereas sensitive attributes are anonymized either using l-diversity or t-closeness or both . Since optimal multidimensional anonymization is an NP-hard problem, in this project an implementation of the Mondrian algorithm [10], which is a greedy approximation algorithm, is used to accomplish this objective.

The greedy algorithm is substantially more efficient when compared to the already existing optimal k-anonymization algorithms as the time complexity of the greedy algorithm is O($n$log$n$), while the optimal algorithms are exponential in the worst case. The greedy multidimensional algorithm often produces higher-quality results than optimal single-dimensional algorithms.
It is assumed that the quasi-identifiers are identified and understood based on specific knowledge of the domain.

Mondrian partition algorithm processes the dataset as follows:
Step 1: The algorithm starts by taking the entire dataset, the quasi-identifying attribute list and the sensitive attribute list as inputs. Considering the entire dataset as a single partition, proceed to step 2.
Step 2: For each partitions, the algorithm then calculates the span of the partition in the following way:

Step 2.1: If the attribute is categorical, then span is set as the number of unique values in that column. *Step 2.2*: Else if the attribute is continuous, then span is set as the difference between the maximum value of the attribute and the minimum value of the attribute within the partition.
Step 3: By taking the span values in decreasing order, we then partition the dataset into two halves (LHS and RHS), each considered as a separate partition.

Step 3.1: If the attribute is categorical, then the split is done by dividing the unique values of the attributes equally across the divisions.

Step 3.2: If the attribute is continuous, then the split is done using the median value.
Step 4: Each of these partitions are then checked separately if it satisfies the defined conditions of l-diversity and/or t-closeness while ensuring that the partition size stays $\geq$ k. If both the partitions does satisfy the conditions, then these successful partitions are then subjected to step 2, where the computation of span is done followed by the attempt to split the partition again.
Step 5: The algorithm will simple loop over if any sub-divided partition fails to satisfy the anonymity conditions, hence not splitting it any further.
Step 6: The algorithm will stop when no partitions can be split any further.

Once these partitions are created, it is clear that these partitions satisfy l-diversity and/or t-closeness and the size of each partition is $\geq k$. Now, the quasi-identifying attributes are generalized appropriately either using generalization hierarchies or simple aggregation measures to make sure that all of these individual partitions satisfy k-anonymity. This is done by replacing all the quasi-identifying attribute values with the same generalized aggregated value for each equivalent classes.

Although the values of k, l and t are given by the analyst, we can propose values and ranges for these parameters as a guide to the analyst. As of now, the proposed tool predicts a value of k by dividing the input datasets into various classes containing same QI values, and taking the average of the size of class with the maximum value and the size of class with the minimum value. The tool also gives the range of l and t values. The range of l is between 1 and the total number of unique values of sensitive attribute and the range of t is from 0 to the overall distribution of the sensitive attribute in the entire dataset.

9.2.3 Insensitive attributes

Insensitive attributes have nothing to do with the individuals present in the dataset, and thus they possess no threat to de-identification. Such attribute values are kept as-is.

9.3 Statistical Measures

Statistical measures like mean, median, mode, variance, standard deviation etc can be derived from a dataset, but the problem with revealing such statistical measures is that this could potentially lead to re-identification of the tuples in the dataset. The inherent nature of the dataset may be biased in such a way that the release of accurate statistical measures can be seen counter-productive to the goal of achieving anonymity. Hence in this work, the statistical measures are published using differential privacy to make sure that adequate noise is added to the computed values to make sure that the resultant values are close to the original values but not precise enough to lead to re-identification.

In this study, we provide differentially private estimates of statistical data measures including mean, median, sum, count, standard deviation, variance, min, and max.

9.4 Utility Measures

Generalized Information Loss (GenILoss) helps us compute the penalty incurred when generalizing a specific attribute, by quantifying the fraction of the domain values that have been generalized. In this study, the normalized version of this metric is used.

Let L_i and U_i be the lower and upper bounds of an attribute i. A cell entry for attribute i is generalized to an interval ij defined by the lower L_{ij} and upper bound U_{ij} end points. The overall information loss of an anonymized table T* can be calculated as:

$$GenILoss(T^*) = \frac{1}{|T| \cdot n} \times \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{U_{ij} - L_{ij}}{U_i - L_i}$$

Figure 9.1: GenILoss Score calculation

where T is the original table, n is the number of attributes and |T| is the number of records.

This metric is based on the concept that attribute values that represent a larger range of values (e.g., ”not married”) are less precise than the ones that represent a smaller range of values (e.g., ”single” or ”divorced”). The output of the GenILoss metric will be in the range [0, 1]. 0 means no transformation (original data) and 1 means full suppression/maximum level of generalization of the data. The analyst should tune the parameters to get the desired output. This metric can be employed in algorithms that uses a generalization hierarchy as well as non-hierarchical algorithms, as any interval of generalized values can be quantified in this way.

Discernibility Metric ($\mathrm{C_{DM}}$) measures how indistinguishable a record is from others, by assigning a penalty to each record, equal to the size of the equivalent class to which it belongs to [11]. If a record is well-suppressed from the other k records in an equivalent class, then each record is assigned a penalty equal to the size of the equivalent class which it belongs to. Otherwise, the record is assigned a penalty equal to the size of the input table. The overall DM score for a k-anonymized table T* is defined by:

$$DM(T^*) = \sum_{\forall EQs.t.|EQ| \geq k} |EQ|^2 + \sum_{\forall EQs.t.|EQ| < k} |T| \cdot |EQ|$$

Figure 9.2: DM Score calculation

where T is the original table, |T| is the total number of records in the table and |EQ| is the size of the equivalence classes created after performing the anonymization. The idea behind this metric is that larger EQs represent more information loss, thus lower values for this metric are desirable.

Average Equivalence Class Size Metric (C_{AVG}) measures how well the creation of the equivalent classes approaches the best case, where the best case being each record generalized in an equivalent class containing k records [11]. A value of $C_{AVG} = 1$ would indicate the ideal anonymization in which the size of the equivalent classes is the given k value. The overall C_{AVG} score for an anonymized table T* is given by:

$$C_{AVG}(T^*) = \frac{|T|}{|EQs| \cdot k}$$

Figure 9.3: C_{AVG} Score calculation

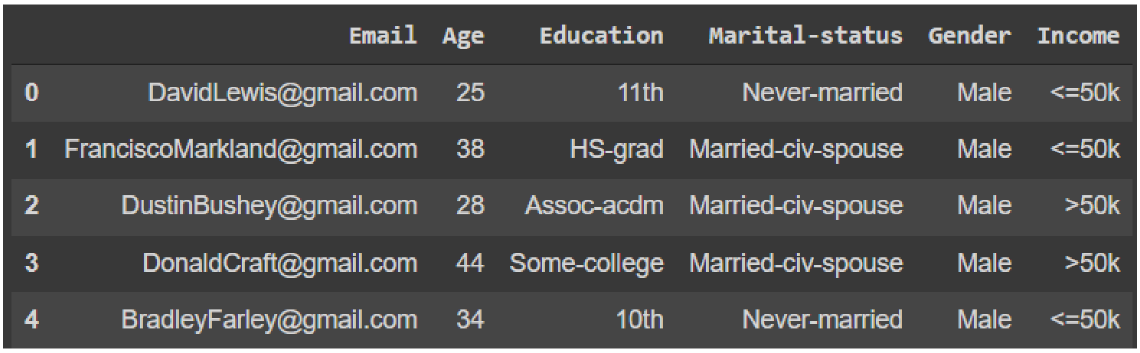
where T is the original table, |T| is the number of records, |EQ| is the total number of equivalence classes created and k is the privacy requirement.

Chapter 10

RESULTS & DISCUSSIONS

In this study, various data anonymization methods like k-anonymity, l-diversity and t-closeness were examined. Also, a specific application of differential privacy was also looked upon.

The dataset used to depict the computations of the anonymization process is a modified version of the US Adult Survey report (https://raw.githubusercontent.com/athulck/Data-Anonymization-Tool/main/adult.sample.csv). This dataset contains 15,060 rows, each having attributes like Email, Age, Education, Marital status, Gender and Income.



	Email	Age	Education	Marital-status	Gender	Income
0	DavidLewis@gmail.com	25	11th	Never-married	Male	<=50k
1	FranciscoMarkland@gmail.com	38	HS-grad	Married-civ-spouse	Male	<=50k
2	DustinBushey@gmail.com	28	Assoc-acdm	Married-civ-spouse	Male	>50k
3	DonaldCraft@gmail.com	44	Some-college	Married-civ-spouse	Male	>50k
4	BradleyFarley@gmail.com	34	10th	Never-married	Male	<=50k

Figure 10.1: Sample Dataset

After making sure that the dataset is parsed properly, we collect the attribute classification information, which tells the tool which all attributes are direct, quasi, sensitive and insensitive.

The tool provides a reasonable prediction for the value of k and a range for l and t. For this dataset, the tool predicted the probable value for k as 42.0, and the range of l as [1, 2] and the range of t as [0.0, 0.75). After this, the user can properly set their values for k, l and t and start the anonymization process. For this test run, we set k=41, l=2 and t=0.

The fig 10.2 depicts a very important characteristic of the dataset. On the x-axis, we have the total number of equivalent classes in the dataset before anonymization. These

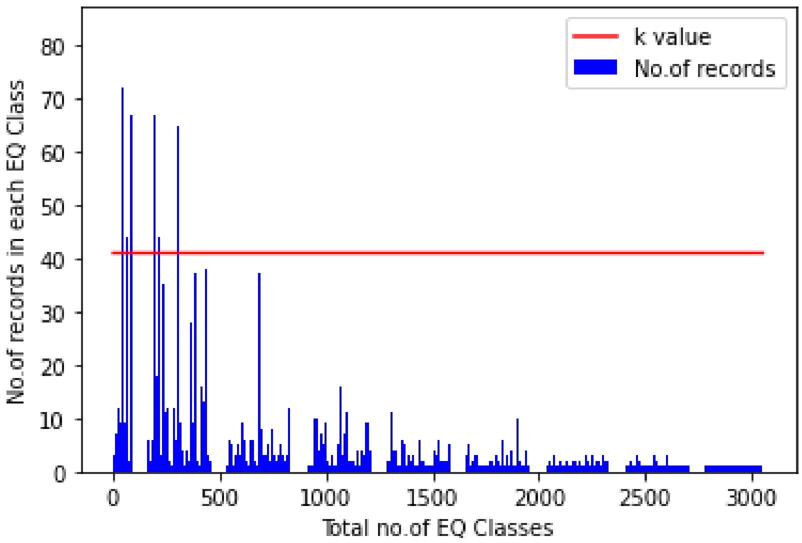
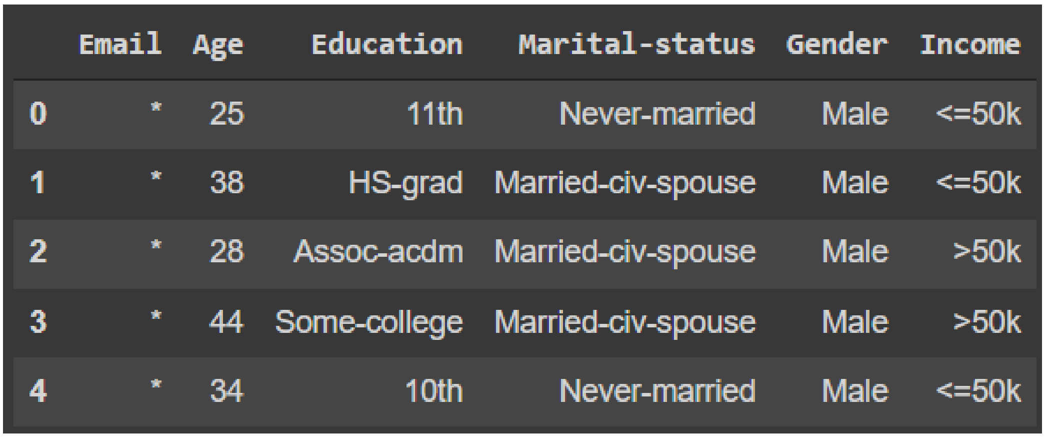


Figure 10.2: Distribution of equivalent classes before anonymization

equivalent classes simply represents the unique set of values for the quasi-identifier attributes. The dataset has 3052 unique combinations of quasi-identifier values and the y-axis represents how many records are there in each of these equivalent classes. For a very common set of quasi-identifier set (like., "(20, 'Some-college', 'Never-married', 'Female')") you can see that there are a lot of records in that specific equivalent class.

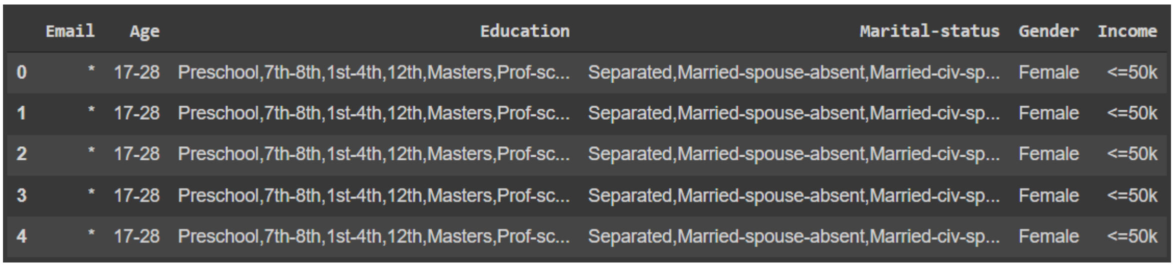
First step in the anonymization process is the suppression of direct-identifiers.



	Email	Age	Education	Marital-status	Gender	Income
0	*	25	11th	Never-married	Male	<=50k
1	*	38	HS-grad	Married-civ-spouse	Male	<=50k
2	*	28	Assoc-acdm	Married-civ-spouse	Male	>50k
3	*	44	Some-college	Married-civ-spouse	Male	>50k
4	*	34	10th	Never-married	Male	<=50k

Figure 10.3: Supression of direct-identifiers with *

The next steps contains the Mondrian algorithm trying to partition the entire dataset into equivalent classes (size $\geq$ k) each satisfying the l-diversity and t-closeness parameters. Once the required partitions are obtained, we proceed to the generalization process where the quasi-identifying attributes in each of these equivalent classes are generalized to make it satisfy k-anonymity.



	Email	Age	Education	Marital-status	Gender	Income
0	*	17-28	Preschool,7th-8th,1st-4th,12th,Masters,Prof-sc...	Separated,Married-spouse-absent,Married-civ-sp...	Female	<=50k
1	*	17-28	Preschool,7th-8th,1st-4th,12th,Masters,Prof-sc...	Separated,Married-spouse-absent,Married-civ-sp...	Female	<=50k
2	*	17-28	Preschool,7th-8th,1st-4th,12th,Masters,Prof-sc...	Separated,Married-spouse-absent,Married-civ-sp...	Female	<=50k
3	*	17-28	Preschool,7th-8th,1st-4th,12th,Masters,Prof-sc...	Separated,Married-spouse-absent,Married-civ-sp...	Female	<=50k
4	*	17-28	Preschool,7th-8th,1st-4th,12th,Masters,Prof-sc...	Separated,Married-spouse-absent,Married-civ-sp...	Female	<=50k

Figure 10.4: A few records of the anonymized dataset

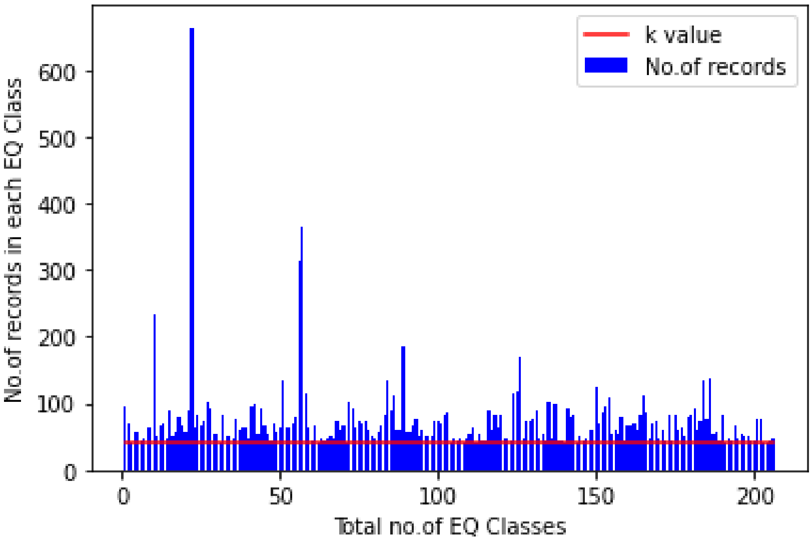


Figure 10.5: Distribution of equivalent classes after anonymization

The resulting dataset is analyzed using the previously defined utility measures.

• Generalization Information Loss (GenILoss) is usually within the range 0 (no transformation) to 1 (full suppression). The anonymized dataset is having the value as 0.0872. Thus, the dataset is generalized as required without much loss in information.

• Discernibility Metric (DM) of the dataset before and after anonymization is measures. Before anonymization it was 176,269,736 and after anonymization it is 1,739,740. Thus we can see a 100X decrease in value. What this means is that more equivalent classes are having its size under k.

- Average Equivalence Class Size Metric beofre anonymization gave a value of 0.120 and after anonymization, the value is 1.783. Although the ideal value is 1, 1.783 is closer to 1 than 0.120. Hence, we can see an improvement.

Chapter 11

FUTURE SCOPE

- ***User group's privacy issue***: Tabular data anonymization mostly focuses on hiding the identity of an individual inside a group of similar individuals. This in-turn reveals the fact that the user belongs to that specific group. Thereby arise the need for user groups' privacy preserving algorithms.

- ***Excessive information loss caused by over-generalization of QIs***: Not all QI's affect the privacy of users in the same way. Sometimes over-generalization of some QIs do cause unwanted information loss without increasing privacy.

- ***Integration of Several Data Sources***: OSINT is never limited to open source data. If the investigator has access to certain closed source data, that can also be fed into the model and it will improve the throughput. Law Enforcement Agencies can use classified data to fine tune the model and expose data points which might be otherwise given less significance. Single data sources can lead to biasing of information and sometimes misinformation. It is necessary to consult multiple data sources in a wide spectrum to understand it to its fullest. The problem here is that the structure of data varies from source to source, and we should define how redundancy in data should be dealt with.

- ***Effective resolution of privacy and utility trade-off***: In data anonymization, there exist a strong trade-off between privacy and utility. Tailoring the anonymization with privacy objectives can adversely affect the anonymous data utility, and vice-versa. This longstanding challenge in the field of PPDP seeking novel solutions to support privacy preserving big data analytics.

- ***Accurate modeling of the adversaries' background knowledge***: Adversaries poses more side channel information as background knowledge (BK) about Social Networks users compared to the tabular data. Recently, text mining and natural language processing (NLP) techniques has increased vastly which further contributes to find and serach on these background data-pool. Thus, accurate modeling of the background knowledge available to an adversary while anonymizing data is very challenging, and further research on how to model the background knowledge during anonymization process is required to effectively protect user's privacy in big data era.

Chapter 12

CONCLUSION

Various researches that has been proposed to anonymize data by upholding utility while preserving user's privacy from malevolent adversaries, namely privacy preserving data publishing (PPDP) techniques, are carefully considered and studied in this work. In recent years, there is an increase in focus on the rapid development of more practical anonymization solutions due to the significant rise in privacy breaches across the globe, and this area is attracting researchers' interests. Although increasing amount of data offers unprecedented opportunities for analytics, but it also introduces challenges in the area of user privacy.

With this work, we analyzed how a tabular data can be anonymized using privacy models such as k-anonymity, l-diversity and t-closeness. The proposed tool also gives predictions on probable values of k and problable range of values for l and t. The statistical measures of the input dataset is anonymized using differential privacy by carefully adding noise into the values. Anonymized data utility is measured using GenILoss, Discernibility metric and Average Equialent Class Size metric so that the data analyst can make carefully thread the balance of data utility and privacy.

The work has explored various models regarding the relational anonymization techniques used for the tabular data. Furthermore, deeper insights on the privacy problems in future computing paradigm were laid out that will be helpful in devising more secure anonymization methods, and we discuss numerous promising open research directions/problems that need further research and developments.

REFERENCES

[1] MeitY. "Report of the Joint Committee on The Personal Data Protection Bill". In: (2021).

[2] Ministry of Health & Family Welfare Government of India. "Digital Information Security in Healthcare, Act (DISHA)". In: (2017).

[3] Abdul Majeed and Sungchang Lee. "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey". In: *IEEE Access* 9 (2021), pp. 8512–8545. DOI: 10.1109/ACCESS.2020.3045700.

[4] Joana Ferreira Marques and Jorge Bernardino. "Analysis of Data Anonymization Techniques". In: (2020). DOI: 10.5220/0010142302350241.

[5] Cristian Augusto et al. "Test-Driven Anonymization in Health Data: A Case Study on Assistive Reproduction". In: (2020), pp. 81–82. DOI: 10.1109/AITEST49225.2020.00019.

[6] Puneet Goswami and Suman Madan. "Privacy preserving data publishing and data anonymization approaches: A review". In: (2017), pp. 139–142. DOI: 10.1109/CCAA.2017.8229787.

[7] Latanya Sweeney. "Simple Demographics Often Identify People Uniquely". In: *Carnegie Mellon University, Data Privacy* (2000). URL: http://dataprivacylab.org/projects/identifiability/.

[8] Maryam Archie et al. "Who's Watching ? De-anonymization of Netflix Reviews using Amazon Reviews". In: (2018).

[9] Working Party European Union. "European Union Working Party on the protection of individuals with regard to the processing of personal data". In: (2014).

[10] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. "Mondrian Multidimensional K-Anonymity". In: (2006), pp. 25–25. DOI: 10.1109/ICDE.2006.101.

[11] Vanessa Ayala-Rivera et al. "A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners". In: *Transactions on Data Privacy* 7 (Dec. 2014), pp. 337–370.

Appendix A

Findings

Personally Identifiable Information
Name
Address
Date of Birth
Telephone Number
Email Address
Password
Financial information such as bank account or credit card or debit card or other payment instrument details
Physical, physiological and mental health condition
Sexual orientation
Medical records and history
Biometric Information
Vehicle number
Any government number, including Aadhar, Voter's Identity, Permanent Account Number ('PAN'), Passport, Ration Card Number

Table A.1: A non-exclusive list of Personally Identifiable Information