An investigation of the application of autoencoders and large-language models to privacy-utility tradeoffs in group-specific settings

by

Bishwas Mandal

B.Tech, Koneru Lakshmaiah Education Foundation, 2019

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Abstract

Machine learning models have emerged as highly effective tools in tackling an array of practical challenges, including image classification, regression tasks, behavioral forecasting, and natural language processing, among others. Their ability to analyze substantial volumes of data endows them with significant value across diverse domains. The widespread success of machine learning models has led data analysts and scientists to extensively incorporate them into predictive and generative modeling endeavors. While leveraging data can enable professionals to gain insights into user preferences and refine predictive and generative accuracy, there exists a risk of inadvertently or intentionally inferring sensitive personal information from the data. Consequently, there is a critical need for the development of privacy mechanisms capable of sanitizing data to shield sensitive attributes while retaining data utility. Achieving this balance is pivotal in safeguarding individual privacy amidst the complexities of the digital age.

In discussions revolving around privacy considerations, there is a prevalent assumption that privacy requirements are homogeneous across the populace, a construct termed as the single-group setting. This dissertation introduces an adversarial learning framework that employs various iterations of autoencoders, specifically designed to navigate the privacy-utility tradeoff within this single-group setting. While the single-group setting holds significance in numerous practical contexts, it fails to consistently capture the varied requirements of different use cases. Recognizing the imperative to address a broader spectrum of scenarios, this dissertation introduces a problem formulation focused on the privacy-utility tradeoff within two distinct user groups. These heterogeneous groups encompass users characterized by differing private and utility attributes, aiming to encompass a wider array of application scenarios within the privacy-utility tradeoff domain.

Furthermore, the landscape of computer science is undergoing significant transformation, particularly in light of the recent surge in large language models (LLMs). LLMs are rapidly evolving, necessitating a thorough examination of the impact of these advancements on privacy domain. Thus, this dissertation endeavors to investigate whether LLMs, endowed with diverse capabilities, can be effectively employed for data sanitization to uphold user privacy. The straightforwardness of this approach without requiring specialized expertise underscores the adaptability of large language models in addressing privacy concerns.

Collectively, this dissertation seeks to enhance the understanding of privacy-utility trade-off in different group specific settings with adversarial learning frameworks and large language models. The findings underscore the potential of these methodologies to provide robust privacy protections while simultaneously preserving or augmenting the utility of data across a myriad of applications.

An investigation of the application of autoencoders and large-language models to privacy-utility tradeoffs in group-specific settings

by

Bishwas Mandal

B.Tech, Koneru Lakshmaiah Education Foundation, 2019

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Approved by:

Major Professor
George Amariucai, PhD

Copyright

Abstract

Machine learning models have emerged as highly effective tools in tackling an array of practical challenges, including image classification, regression tasks, behavioral forecasting, and natural language processing, among others. Their ability to analyze substantial volumes of data endows them with significant value across diverse domains. The widespread success of machine learning models has led data analysts and scientists to extensively incorporate them into predictive and generative modeling endeavors. While leveraging data can enable professionals to gain insights into user preferences and refine predictive and generative accuracy, there exists a risk of inadvertently or intentionally inferring sensitive personal information from the data. Consequently, there is a critical need for the development of privacy mechanisms capable of sanitizing data to shield sensitive attributes while retaining data utility. Achieving this balance is pivotal in safeguarding individual privacy amidst the complexities of the digital age.

In discussions revolving around privacy considerations, there is a prevalent assumption that privacy requirements are homogeneous across the populace, a construct termed as the single-group setting. This dissertation introduces an adversarial learning framework that employs various iterations of autoencoders, specifically designed to navigate the privacy-utility tradeoff within this single-group setting. While the single-group setting holds significance in numerous practical contexts, it fails to consistently capture the varied requirements of different use cases. Recognizing the imperative to address a broader spectrum of scenarios, this dissertation introduces a problem formulation focused on the privacy-utility tradeoff within two distinct user groups. These heterogeneous groups encompass users characterized by differing private and utility attributes, aiming to encompass a wider array of application scenarios within the privacy-utility tradeoff domain.

Furthermore, the landscape of computer science is undergoing significant transformation, particularly in light of the recent surge in large language models (LLMs). LLMs are rapidly evolving, necessitating a thorough examination of the impact of these advancements on privacy domain. Thus, this dissertation endeavors to investigate whether LLMs, endowed with diverse capabilities, can be effectively employed for data sanitization to uphold user privacy. The straightforwardness of this approach without requiring specialized expertise underscores the adaptability of large language models in addressing privacy concerns.

Collectively, this dissertation seeks to enhance the understanding of privacy-utility trade-off in different group specific settings with adversarial learning frameworks and large language models. The findings underscore the potential of these methodologies to provide robust privacy protections while simultaneously preserving or augmenting the utility of data across a myriad of applications.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I extend my heartfelt appreciation to my advisor, George Amariucai, for his steadfast guidance and support throughout my PhD journey. His invaluable ideas and insights have greatly contributed to the success of this dissertation, making my doctoral experience both enriching and rewarding. Additionally, I am grateful to Shuangqing Wei from Louisiana State University, who co-advised my research and provided invaluable insights that significantly enhanced its quality. Together, George and Shuangqing made my PhD journey not only educational but also enjoyable and fulfilling.

I would also like to thank my PhD committee members, Doina Caragea, Torben Amtoft, and Bala Natarajan, for their invaluable mentorship and constructive feedback. Their guidance has steered me towards promising research directions and has undoubtedly played a pivotal role in my development as a researcher.

I extend my gratitude to the Department of Computer Science for their financial support, which facilitated access to large language models' API access and sponsored travel expenses for presenting research findings at various conferences.

A special mention is reserved for Chandra Sharma, who provided invaluable support from the moment I arrived in the United States. His assistance, including picking me up at the airport to having insightful discussions, collaborations and meaningful conversations, greatly influenced my research endeavors. I am also grateful to Bipin Paudel for engaging in meaningful discussions on research topics, which aided in identifying and addressing research challenges.

I am also thankful to my other lab freinds, Adaeze Okeukwu and Stephanie Harshbarger, who not only enriched my PhD experience but also made the lab feel like home with their friendship and support.

Last but certainly not least, I owe a debt of gratitude to my family. To my father, Jeebachh Mandal, my mother, Sushila Kumari Mandal, and my brother, Vishal Mandal, who have been my pillars of strength throughout my life. My father's inspiration and my brother's unwavering support have been instrumental in my journey. Finally, I am profoundly thankful to my wife, Salina Dahal, for her constant love, support, and encouragement. Her unwavering presence has been my anchor, guiding me through challenges and celebrating my successes. She is truly everything I could have ever hoped for.

Dedication

My dear parents and brother,

Whose enduring love, unwavering support, and heartfelt blessings have been my guiding force throughout this journey.

In loving memory of my grandfather,

Whose memory and encouragement have always inspired me to reach new heights.

And to my beloved wife,

For her boundless love, constant encouragement, and unwavering presence through every step of this endeavor.

Chapter 1

Introduction

Over the past decade, there has been a remarkable surge in the utilization of social media platforms. These platforms serve as avenues for individuals to articulate their viewpoints, share information, insights, and various other forms of content. Users frequently disseminate their moments of happiness, sadness or apprehension through diverse media formats including photos, videos, audio, etc. Other different form of interactions occurring on social media are stored in relational databases generating standard tabular datasets (Shwartz-Ziv and Armon, 2021). Consequently, this surge in user-generated content has resulted in an unparalleled volume of data. This influx of data presents both opportunities and challenges for understanding human behavior and societal trends in the digital age.

Machine learning (deep learning), a field that thrives on large datasets, has experienced remarkable success in recent years. The rapid expansion witnessed in machine learning can be attributed to several factors, including advancements in architectural designs, enhanced convergence capabilities, and the widespread availability of open-source models. Moreover, the increasing integration of machine learning into diverse industries and the growing emphasis on AI research have contributed to its widespread adoption and continuous evolution. However, the primary driving force behind the substantial growth of machine learning remains the accessibility of vast datasets and the continual improvement of computational

resources (Sarker, 2021).

The development within the realms of machine learning models has enabled them to demonstrate outstanding efficiency in confronting a diverse range of practical challenges, encompassing tasks such as image classification, regression analysis, prediction of human behavior, natural language processing, and an array of other applications. The datasets necessary for training these models can contain significant amounts of sensitive information. The sensitivity of the data is contingent upon the application domain and the perceptions of users. In some instances, the entirety of the data may be sensitive, such as when dealing with credit card numbers, health records, private conversations, or images. Alternatively, even if the data itself is not inherently private, the information that can be inferred from it, due to correlations, may still be sensitive (Sun et al., 2018; Sun and Tay, 2020).

For instance, when the data itself is private, attackers may employ various techniques, such as model inversion attacks or membership inference attacks, etc. with the aim of reverse engineering the original data used to train machine learning (ML) models or determining whether a specific data point was included in the training dataset. Traditional methods for preserving privacy include Differential Privacy (DP) (Dwork and Roth, 2014), Homomorphic Encryption (HE) (Acar et al., 2018), secure multi-party computation (SMPC) (Evans et al., 2018). DP primarily ensures that databases with a single differing record remain indistinguishable, primarily serving as a defense against membership inference attacks. HE enables computations to be performed on encrypted data, while SMPC allows data to remain on users' devices, thereby preserving data privacy by preventing it from leaving the user's device.

Similarly, in cases where the data itself may not be private but the extracted information from the data might be sensitive, individuals' privacy may be compromised, or they could be subject to discrimination. For instance, in scenarios involving facial recognition technologies capable of identifying health conditions, while such capabilities hold value in certain healthcare applications, malicious actors armed with sufficient data and labels could

construct similar models to infer health condition of users. Although the facial images per se may not be sensitive, the potential invasion of privacy by such technology poses a significant risk. Hence, in such scenarios, it becomes imperative to perturb the data in a manner that prevents the extraction of sensitive features by models developed for nefarious purposes. Similarly, machine learning (ML) models often exhibit various biases, as evidenced by the COMPAS data revealing algorithmic bias against African American defendants (Khademi and Honavar, 2019). Attributes like race, religion, national origin, gender, marital status, age, socioeconomic status, and disability status are commonly regarded as sensitive and can serve as sources of potential bias in various contexts, including machine learning and decision-making systems. Although models may not explicitly aim to infer such information, their inadvertent biases can lead to unfair outcomes for certain population groups.

This dissertation endeavors to address privacy concerns in scenarios where the data itself might not be private, but sensitive information about individuals can be inferred from it. Various iterations of this classical privacy problem exist, with some variants involving scenarios where utility is typically measured using a distortion function (Sankar et al., 2013b; Makhdoumi and Fawaz, 2013a; Rajagopalan et al., 2011a; Erdogdu and Fawaz, 2015), which assess the overall loss of information between the original and perturbed data, while privacy pertains to the inability to deduce private features. In the realm of machine learning, however, utility typically refers to a classifier's ability to accurately classify any utility feature while failing to accurately classify the private feature, or reducing the bias in decision-making processes when classifying the utility feature. To address the challenges posed by the trade-off between privacy and utility, diverse adversarial learning frameworks utilizing various forms of autoencoders are implemented (Edwards and Storkey, 2016; Chen et al., 2018, 2019). Consequently, this dissertation aims to investigate different autoencoder variants used in privacy mechanisms, and the use of large language models, specifically GPT-4, capable of enabling machine learning models to accurately infer the utility feature while preventing them from accurately classifying the private feature within specific group settings.

1.1 Contributions of this dissertation

The prevalence of social media usage has resulted in widespread sharing of personal data by nearly all individuals. Yet, due to a lack of awareness among regular users regarding the privacy implications of their actions, numerous privacy concerns persist. Hence, in this dissertation, we propose the following approaches aimed at addressing diverse privacy related issues.

1.1.1 Uncertainty-Autoencoder-Based Privacy and Utility Preserving Data Type Conscious Transformation

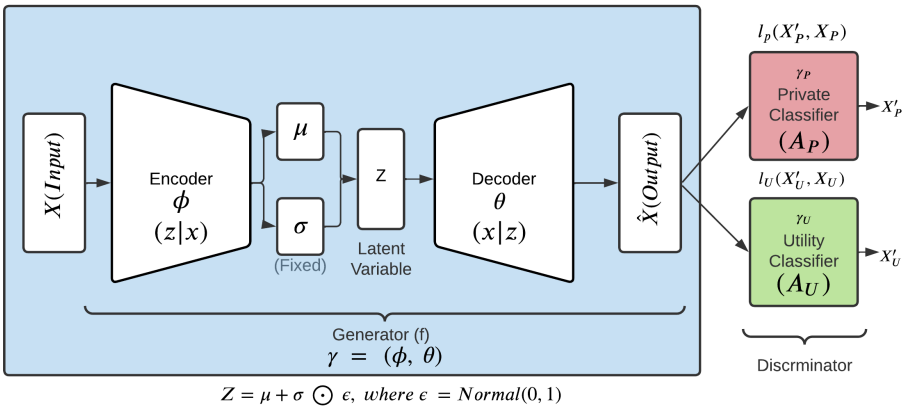


Figure 1.1: Overview of the training process of Uncertainty Autoencoder based privacy mechanism.

When addressing privacy mechanisms, existing works typically focus either on representations or, when considering the final sanitized data, often overlook the categorical nature of the data itself. This work introduces a privacy preserving Uncertainty Autoencoder designed to navigate the privacy-utility tradeoff problem under two distinct conditions: data-type ignorant and data-type aware. Under data-type aware conditions, the privacy mechanism provides one-hot encoding of categorical features, representing exactly one class. Conversely, under data-type ignorant conditions, the categorical variables are represented by a collec-

tion of scores, one for each class. The Uncertainty Autoencoder, trained adversarially, is a variant of the autoencoder architecture comprising an encoder-decoder pair. Notably, during training, as shown in Figure 1.1, alongside the Uncertainty Autoencoder (generator), a discriminator incorporating an adversary and a utility provider is employed. Unlike prior research utilizing autoencoders (AEs) without introducing randomness, or variational autoencoders (VAEs) (Kingma and Welling, 2014) relying on learned latent representations constrained to a Gaussian assumption, the Uncertainty Autoencoder introduces randomness and eliminates the Gaussian assumption on latent variables, emphasizing an end-to-end stochastic mapping from input to sanitized data.

1.1.2 Optimizing Privacy and Utility Tradeoffs for Group Interests Through Harmonization

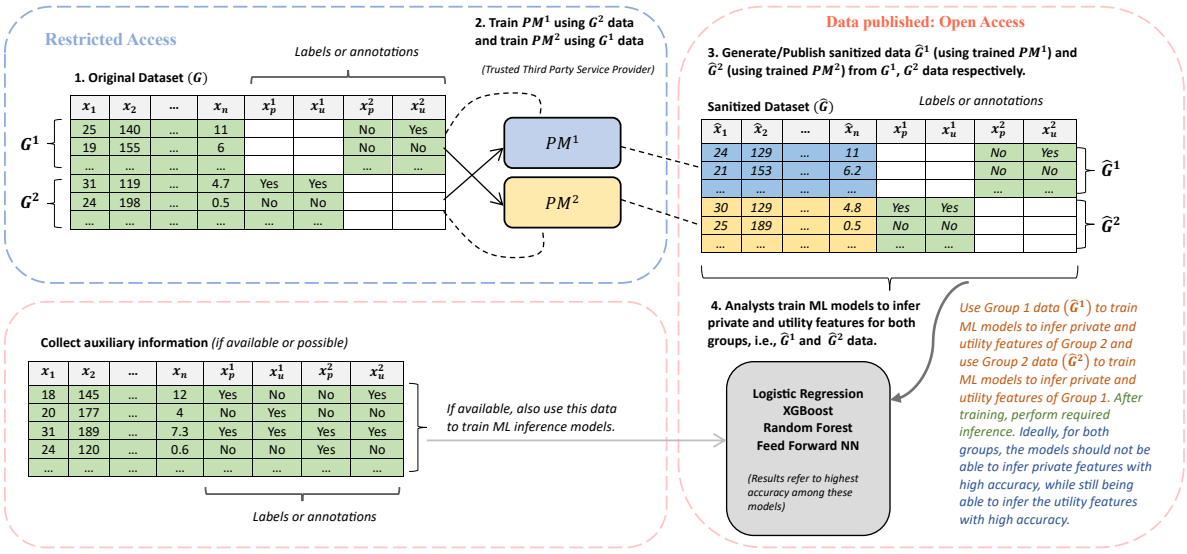


Figure 1.2: Overview of the privacy-utility tradeoff in two-user group setting.

This work introduces an innovative problem formulation aimed at managing the privacy-utility tradeoff within contexts involving two disparate user groups distinguished by distinct private and utility attributes, as shown in Figure 1.2. In contrast to prior research, which

predominantly concentrates on scenarios where all users possess identical private and utility attributes and frequently depends on auxiliary datasets or manual annotations, this project proposes a collaborative data-sharing framework facilitated by a trusted third party. This intermediary employs adversarial privacy techniques within the data-sharing mechanism to internally sanitize data for both user groups, thereby obviating the necessity for manual annotation or auxiliary datasets to navigate the privacy-utility tradeoff problem.

1.1.3 Privatization Mechanisms to Attain Privacy and Utility Tradeoffs in Multi-Party Environments

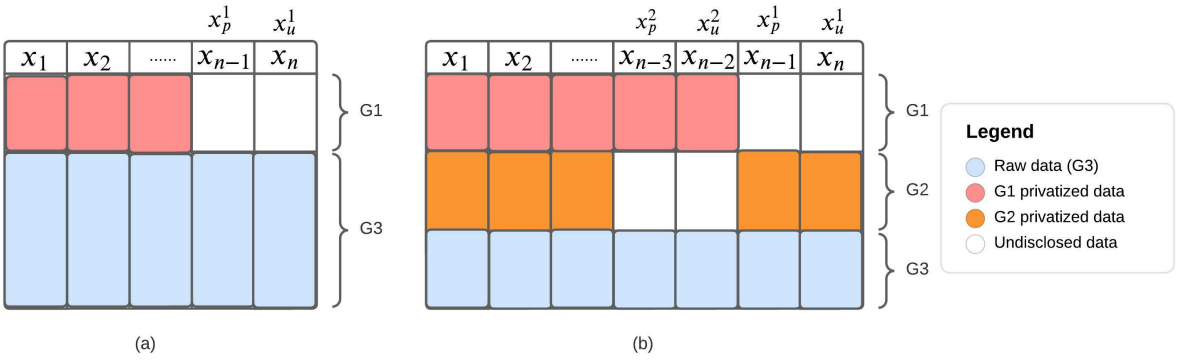


Figure 1.3: Single Party Privacy Utility Tradeoff setting versus Multi-Party Privacy Utility Tradeoff setting.

This work is a continuation of previous efforts, wherein we referred to a scenario involving two groups. Here, multiple parties denotes various groups of users. Despite this broader context, we focus on a simple scenario involving only two parties necessitating privacy measures. However, in our earlier work, a trusted third-party service provider was present, enjoying the trust of both user groups and thereby receiving original data from both. Within this framework, training privacy mechanisms was comparatively straightforward since the third party possessed all clean data and labels. In this work, we extend this previous work by removing the assumption of a trusted third-party service provider. Consequently, the

data disclosed in the final database comprises not only perturbed data but also perturbed labels, as shown in Figure 1.3. We introduce three training mechanisms to augment the existing privacy mechanisms, addressing this new challenge.

1.1.4 Initial Exploration of Zero-Shot Privacy Utility Tradeoffs in Tabular Data Using GPT-4

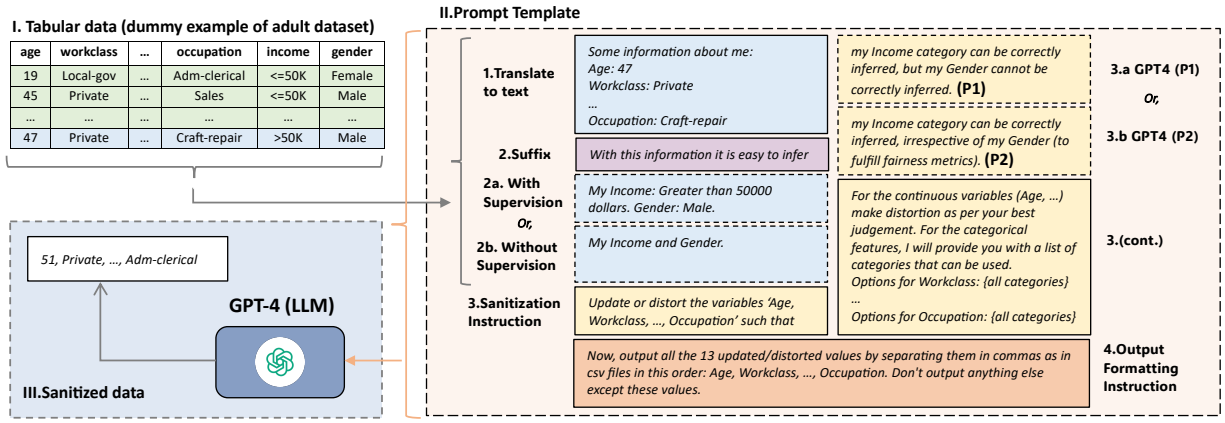


Figure 1.4: Overview of the proposed data sanitization technique using GPT-4. This approach involves initializing the GPT-4 model with a prompt that includes the conversion of tabular data points into textual format, followed by the incorporation of sanitization instructions.

The privacy implications of Large Language Models (LLMs) encompass various concerns. While existing research predominantly focuses on extracting training data from LLMs, this is not the sole privacy issue they present. The increasing inference capabilities of LLMs have highlighted that their privacy concerns extend beyond mere memorization of training data (Staab et al., 2023). Given the demonstrated excellence in inference capabilities of LLMs, particularly GPT-4, this work explores their application in scenarios involving the tradeoff between privacy and utility in tabular data. However, instead of assessing vulnerability to privacy concerns, we utilize LLMs as a defense mechanism. This work, depicted in Figure 1.4, involves prompting GPT-4 by converting tabular data points into textual format and providing precise sanitization instructions in a zero-shot manner. The

primary goal is to sanitize tabular data to impede existing machine learning models from accurately inferring private features while enabling accurate inference of utility-related attributes. This relatively straightforward approach yields performance comparable to more complex adversarial optimization methods used for managing privacy-utility tradeoffs.

1.1.5 Summary

In summary, this dissertation makes contributions to the investigation of strategies directed at effectively managing the tradeoffs between privacy and utility. It addresses privacy concerns in scenarios where the data itself might not be sensitive but the information that can be inferred from the data is sensitive. The contributions of this work have already been *published/accepted* in peer-reviewed conferences, as shown below:

- **B. Mandal**, G. Amariucai and S. Wei, "Uncertainty-Autoencoder-Based Privacy and Utility Preserving Data Type Conscious Transformation," 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022, pp. 1-8, doi: 10.1109/IJCNN55064.2022.9892789. URL: https://ieeexplore.ieee.org/abstract/document/9892789

- **B. Mandal**, G. Amariucai and S. Wei, "Optimizing Privacy and Utility Tradeoffs for Group Interests Through Harmonization," 2024 IEEE International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 2024 *(accepted)*.

- **B. Mandal**, G. Amariucai and S. Wei, "Initial Exploration of Zero-Shot Privacy Utility Tradeoffs in Tabular Data Using GPT-4," 2024 IEEE International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 2024 *(accepted)*.

In addition to the publications directly integrated into this dissertation, the following peer-reviewed publications represent supplementary contributions that I have made throughout my doctoral studies.

- C. Sharma, **B. Mandal** and G. Amariucai, "A Practical Approach to Navigating the Tradeoff Between Privacy and Precise Utility," ICC 2021 - IEEE International Conference on Communications, Montreal, QC, Canada, 2021, pp. 1-6, doi: 10.1109/ICC42927.2021.9500410. URL: https://ieeexplore.ieee.org/abstract/document/9500410

- **Bishwas Mandal**, Sarthak Khanal, and Doina Caragea. 2024. Contrastive Learning for Multimodal Classification of Crisis related Tweets. In Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3589334.3648143 *(accepted)*.

Chapter 2

Uncertainty-Autoencoder-Based Privacy and Utility Preserving Data Type Conscious Transformation

2.1 Introduction

Amidst the recent surge in online service utilization, users increasingly share substantial volumes of data with various service providers in exchange for utility. These providers leverage diverse machine learning models, some aimed at enhancing user experience and others geared towards business promotion. In efforts to understand their user base or promote their services, providers may deploy models capable of inferring multifaceted user information. Despite users' diligence in safeguarding what they perceive as private data, including identity, age, race, location, gender, income, medical history, and political affiliations, the shared data may still contain a significant amount of information pertaining to these attributes. This abundance of data poses a risk, potentially enabling intruders/third-party service providers to infer sensitive user information. Notably, Google recently terminated its FLoC (Federated Learning of Cohorts) initiative following concerns regarding the potential

inference of sensitive user data from behavioral patterns and interests. It is therefore crucial to recognize these potential data correlations and implement privacy mechanisms capable of obfuscating the link between shared and private data, while retaining essential information necessary to fulfill desired utility levels. We call such a mechanism a *privacy and utility preserving end-to-end transformation (PUPET)*.

Ideally, a PUPET should minimize the leakage of information about the private attributes, and maximize the information that the shared data contains about the utility attributes. The operation of a privacy mechanism is said to achieve a privacy-utility tradeoff when it establishes certain points in some private-information-utility-information plane. Most notions of privacy, such as Differential Privacy (Dwork, 2006; Dwork and Roth, 2014), or k-anonymity (Sweeney, 2002) are achieved by using some sort of distortion, such as adding random noise to the data, or performing data suppression or generalization. The privacy-utility tradeoff is the subject of a large body of research, with many works focused on producing application-specific solutions for the problem (Rajagopalan et al., 2011b; Alvim et al., 2011; Makhdoumi and Fawaz, 2013b; Sankar et al., 2013a; Sharma et al., 2021). Domingo-Ferrer and Torra (2008) shows the drawbacks of k-anonymity and its variant. Similarly, it is known that computing the optimal noise addition in higher dimensional data for differential privacy can potentially be infeasible. Therefore, a wide majority of recent techniques use neural networks and adversarial learning, which can deal with the high dimensional data and can provide an approximation of the underlying functions. In particular, research works such as Edwards and Storkey (2016); Huang et al. (2017); Madras et al. (2018); Pittaluga et al. (2018); Osia et al. (2020); Huang et al. (2019); Chen et al. (2019); Wu et al. (2020); Morales et al. (2020); Xiao et al. (2020); Erdemir et al. (2021); Wu et al. (2021) have been carried out leveraging autoencoder (AE) or variational autoencoder (VAE) (Kingma and Welling, 2014) and/or generative adversarial network (GAN) type training (Goodfellow et al., 2014) or some other adversarial optimization techniques such as GRL (Ganin and Lempitsky, 2015) and K-Beam (Hamm and Noh, 2018). Simi-

larly, Chen et al. (2018) utilized the loss function of both VAEs and GANs together along with their application specific classifier. AEs or VAEs are used to create compressed latent representations capturing minimum and maximum information about the private and utility features, respectively. AEs perform compression alone, without introducing randomness, while VAEs introduce randomness when sampling from the mean and variance of the latent space – they also model explicitly the prior distribution over the latent variable. VAE loss functions include a regularization term to minimize the KL divergence between the variational posterior over the latent variable and some prior distribution (which, for the convenience of generating data in the absence of an encoder, is usually chosen to be white Gaussian). GANs, on the other hand, involve a min-max game between the generator and discriminator. Similarly, obfuscation mechanisms Raval et al. (2019); Ilanchezian et al. (2019); Hsu et al. (2020) demonstrate privacy preservation by first estimating each feature's information density relative to the private data, and then selectively distorting the most information-leaking features.

All these different techniques require either training a filter that adds noise, or adding noise directly into the input stream while training, or introducing distortion via lossy compression or some other form of masking technique. In addition, most of these techniques are tested on image datasets, for which there is no restriction on the output of data, compared to categorical features, where the output should be a one-hot encoding that represents precisely one class. In the adversarial learning framework, most research involving datasets that include categorical features either uses an adversary trained directly on the latent variables (Edwards and Storkey, 2016; Chen et al., 2019), or one trained on the generator output but without enforcing any constraints based on the data type (data-type-ignorant conditions) (Madras et al., 2018). This lack of constraints mean that the output of the privacy mechanism doesn't represent an exact class, but rather some scores associated with different classes. The notion of adding noise or distortion to categorical variables in data publishing can be impractical as users are likely to publish data at a higher level where they

will be limited to mentioning one particular class. Another notable issue in existing works is that most of the privacy mechanisms are tested under a single and particular type of adversary network which cannot demonstrate the *actual privacy leakage* as there might be different adversary models capable of inferring private features with higher accuracy which can diminish the claims of particular privacy mechanisms. For example, Raval et al. (2017) proposed *strong adversary* – an adversary that is trained using the perturbed data generated by their privacy mechanism and ground truth labels – and show that the adversary's accuracy increased from 50% (weak adversary – trained using ground truth image and labels) to 75% i.e. a 25% (absolute) increase in inferring private information. In order to make the privacy mechanism robust against multiple adversaries, Wu et al. (2020, 2021) proposed *privacy budget model restarting and ensemble*.

Inspired by these studies, this work proposes an adversarial learning framework that deals with the privacy-utility tradeoff problem under two conditions: data-type-ignorant, and data-type-aware conditions. Data-type-ignorant refers to the classic privacy-utility tradeoff setting where the categorical variables after privatization are not enforced to belong to one particular class. The data-type-aware condition, on the other hand, refers to the optimization settings where the output of the *PUPET* generator enforces the correct representation of categorical variables by assigning a one to the *argmax* index of the probability distribution and assigning zero to all other indices before passing the data to the adversary and utility provider in the *testing phase*. Thereafter, the privacy mechanism proposed in this study is tested on multiple adversaries– some trained from ground truth data and true labels (weak adversaries) and some trained from the perturbed data generated from the privacy mechanism and true labels (strong adversaries). Furthermore, in the proposed privacy mechanism, we introduce randomness and remove the Gaussian assumption on the prior distribution of the latent variable and only focus on the end-to-end stochastic mapping to transform data that preserves privacy and utility. The Gaussian assumption used by VAEs is useful for ancestral sampling, which is not required in the generation of privatized data. Instead, a

Markov chain ($X \to Z \to \hat{X}$) is required where X is the input data, Z is the latent variable and $\hat{X}$ is the privatized data. To implement the privacy mechanism, Uncertainty Autoencoders (UAEs) (Grover and Ermon, 2019) is utilized, which defines an implicit generative model without specifying a prior on the latent representation. The use of a Markov chain to generate privatized data is necessary regardless of the generative model (AE, VAE, or UAE) and thus, the use of UAE doesn't make the process any more computationally expensive than previous approaches. In our adversarial setting, UAE behaves as a generator and learns from its own loss, along with the loss of the discriminators, to generate privatized data that maintains its utility to the utility provider. On the other hand, the discriminators (adversary and utility provider) learn to infer private and utility features, respectively, from the privatized data generated by the generator.

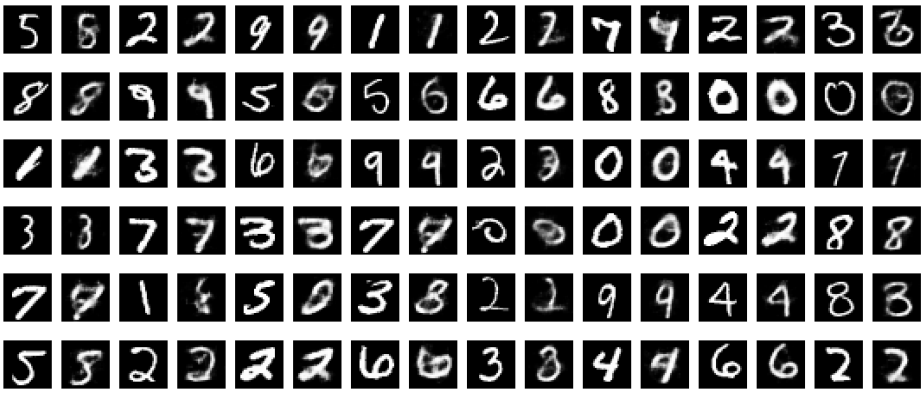


Figure 2.1: **MNIST**: Odd and even adjacent columns show original and privatized versions respectively. For most images, numbers are still in the same category (utility attribute: ≥ 5 or < 5) while being switched from odd to even (private attribute). Some digits change from odd to even but also switch from ≥ 5 to < 5, and some remain unchanged.

To showcase the effectiveness of the proposed privacy mechanism, comprehensive experiments on four datasets viz. MNIST handwritten digits (LeCun et al., 2010), Fashion MNIST (Xiao et al., 2017), UCI Adult (Dua and Graff, 2017), and US Census Demographic Data (US Census Bureau, 2018) is performed. Except UCI Adult, the other datasets do not contain categorical features and thus fit well with data-type-ignorant conditions. Firstly, we

evaluate the proposed method under data-type-ignorant conditions for all the four datasets considering multiple adversary models and compare results with the existing mechanisms. In particular, we directly compare our proposed UAE-based PUPET (or *UAE-PUPET*) mechanism to that of emb-g-filter (Chen et al., 2019), in the same setting as in Chen et al. (2019), using their MNIST Case 2 (variant). We show that our mechanism attains 2.6% less accuracy for the private feature, and 11.2% higher accuracy for the utility feature, thus clearly outperforming the previously existing mechanism. It is worth mentioning that this result was obtained by UAE-PUPET under the best-performing adversary out of multiple adversaries that includes weak adversary, strong adversary, and adversary trained together with utility provider and privacy mechanisms, at the time of adversarially training the privacy mechanism. This shows that our mechanism *UAE-PUPET* is fairly robust, despite the use of a single-adversary model at the time of training, instead of an ensemble of adversaries as in Wu et al. (2020) and Wu et al. (2021), and also provides better privacy and utility guarantees than Chen et al. (2019), which is tested against a single adversary model. Thereafter, we evaluate our privacy mechanism under data-type-aware conditions for the UCI Adult dataset and demonstrate similarity in performance between the claimed leakage (data-type-ignorant condition when dataset has categorical features) and actual leakage in practical applications.

2.2 Problem Formulation and Methodology

Consider a setting where a user wishes to release some data vector X with the intent to receive certain level utility, while maintaining a certain level of privacy about a specific feature or set of features. We represent the private feature vector as X_P and the utility feature vector as X_U, and expect that they are both correlated with X. To ensure the desired privacy and utility guarantees, before publicly sharing their data X, users employ a PUPET that takes X as input, and generates $\hat{X}$, a distorted version of X which contains

minimum information about X_P and maximum information about X_U. The data $\hat{X}$ is then shared publicly.

It is very important to note here that the disclosed data $\hat{X}$ is usually required to preserve in general the size and structure of X. That is, while it may be tempting to devise a compression mechanism – for example, through a simple arbitrary affine transformation – that produces a shorter $\hat{X}$, each component of which is some affine transformation of the components of X, such a mechanism has very little applicability in practice. This is because usually the disclosure of $\hat{X}$ has to take place over a pre-existing platform, outside of the data owner's control, which is designed specifically for X. This platform usually has very specific fields that the user needs to fill out, so that the structure of X is enforced on the PUPET's output. It is also worth noting that most currently existing platforms (such as social media platforms) would provide their utility by taking the values of $\hat{X}$ at face value, as if it was the original X that was being disclosed, which motivates some of the existing privacy-utility tradeoff works to use the distance between $\hat{X}$ and X as a measure of utility (Asoodeh et al., 2015; Erdogdu and Fawaz, 2015; Wang and Calmon, 2017; Kalantari et al., 2017; Basciftci et al., 2016; Wang et al., 2018; Rassouli and Gündüz, 2019; Diaz et al., 2019). However, unlike these works, we consider a smart and informed utility provider, who is aware of the privacy mechanism employed by the user – of course, without knowing the exact realizations of the randomness it uses, and can make a competent inference about the utility feature, using a neural-network-based architecture.

This setup raises questions as to who would use such services in real life, as well as whether such a setup would actually occur. One of the many examples is the creation of an e-commerce account, when certain information would need to be entered. The underlying platform gathers the information we share, and may develop implicit models to recognize, e.g., our gender and income. However, users might not be comfortable knowing that these implicit models learn about their income, but at the same time would also want to get better recommendation of products based on their gender. In such a case, the users can

use a PUPET which generates $\hat{X}$, and then use $\hat{X}$ to fill up the information to create an account. The training of the PUPET requires multiple tuples (X, X_P, X_U), collected from users who do not mind disclosing X_P, X_U, and can be handled either by the data owner, or by a trusted service provider. Similarly, social media giants such as Facebook, Instagram, Twitter etc. sell users' data to different vendors. Such platforms can also use the formulated privatization services by ensuring the vendor only receives the data they paid for, i.e. privatize all the information about other features which they don't get paid for. Countless other applications can be envisioned.

Formally, denote $X^j = \{x_1^j, x_2^j, x_3^j, \cdots x_{n^j}^j\}$, where the components x_i^j are all correlated random variables denoting n distinct features of the user U^j, which the user wishes to release to the public. In addition, the user U^j contains private features X_P^j, and utility features X_U^j, with n_p^j and n_u^j component random variables, respectively. Note that X_P^j and X_U^j are both correlated with X^j and no private and utility feature is in X^j i.e. $X_P^j \notin X^j$ and $X_U^j \notin X^j$. For different users, the choice of the data they wish to share, as well as the private features and the utility features, differ and thus our privacy mechanism needs to be trained differently for different sets of users. For simplicity we drop the user-specific indices and refer to the random vectors directly as X, X_P and X_U. We represent the PUPET in its most general form as a function f (which could be a randomized mapping) that takes input (X, X_P, X_U) and generates $\hat{X}$ i.e. $\hat{X} = f(X, X_P, X_U)$. The adversary builds a learning algorithm a_p that takes privatized data $\hat{X}$ to infer the private attributes X'_P which is an estimate of X_P i.e. $X'_P = a_p(\hat{X})$. The goal of adversary is to minimize the loss between X'_P and X_P i.e. $l_P(X'_P, X_P) = l_P(a_p(f(X, X_P, X_U)), X_P)$. Correspondingly, the utility provider builds a learning algorithm a_u to infer X'_U which is an estimate of X_U and desires to minimize the loss $l_U(X'_U, X_U) = l_U(a_u(f(X, X_P, X_U)), X_U)$. The privacy mechanism f is now chosen to maximize the inference loss l_P and minimize the inference loss l_U. This setting refers to the min-max game expressed as follows:

$$\max_{f \in \mathcal{F}} \Bigg\{ \lambda_P \min_{a_p \in \mathcal{A}_p} \mathbb{E}\left[l_P\left(a_p\left(f\left(X, X_P, X_U \right) \right), X_P \right) \right] \\ -\lambda_U \min_{a_u \in \mathcal{A}_u} \mathbb{E}\left[l_U\left(a_u\left(f\left(X, X_P, X_U \right) \right), X_U \right) \right] \Bigg\}, \tag{2.1}$$

where λ_P, λ_U are hyperparameters that control the tradeoff between adversary loss and utility provider loss, and the expectation is taken over all samples of the dataset, and $\mathcal{F}$, $\mathcal{A}_p$ and $\mathcal{A}_u$ are the sets of functions from which the privacy mechanism, the adversary and the utility provider select their corresponding operators, respectively. In this work we utilize neural networks to optimize over different functions f, a_p, a_u and the losses l_P and l_U are instantiated as the cross-entropy loss (de Boer et al., 2004).

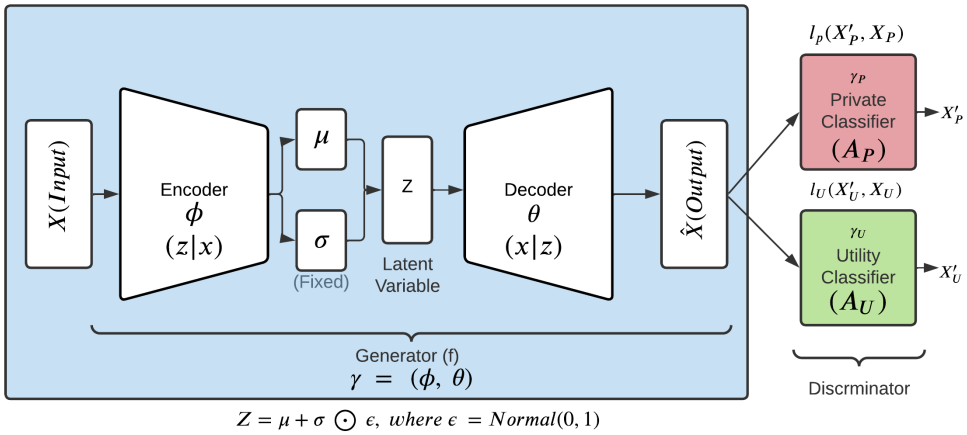


Figure 2.2: **UAE-PUPET** architecture. This architecture supports both conditions (data-type-ignorant and aware conditions). After the completion of training, we detach the discriminator, and use the generator to generate privatized data $\hat{X}$.

In order to solve the optimization problem in Eq.(2.1), we leverage an uncertainty autoencoder (UAE), which serves as the generator for our privacy mechanism. The objective of UAE is given by $\max_{\theta,\phi} \mathbb{E}_{Q_\phi(X,Z)}\left[\log p_\theta(x|z)\right]$, where X is the input data distribution, Z is the latent variable, ϕ, θ are parameters of encoder and decoder, respectively, $Q_\phi(X, Z)$

is the true joint distribution of the input and latent representation, and $p_\theta(X|Z)$ is the likelihood function produced by the decoder, which aims to emulate $Q_\phi(X|Z)$ as closely as possible. Notice that UAE does not force a Gaussian assumption, or any other distribution assumption, on the latent variable marginal distribution, and thus we don't have a KL divergence term in the objective of the UAE, as is the case for the VAE. Instead we focus only on the end-to-end stochastic mapping. Additionally, UAE's encoder outputs only the mean, and sampling is done in the latent space with a fixed variance. If the variance of the noise sampled in the latent space equals zero, the learning objective of the UAE is the same as that of a standard AE. In our adversarial setting, we introduce the parameter γ, which represents the weights of the generator (encoder-decoder pair) with function f (basically, $\gamma = (\phi, \theta)$). Additionally, the output of the generator is attached to the discriminator, which consists of an adversary and a utility provider as shown in Figure 2.2. The adversary learns the function a_p with parameters γ_P to minimize the privacy-specific loss $l_P(X'_P, X_P)$. Similarly, the utility provider learns the a_u with parameters γ_U to minimize the utility-specific loss $l_U(X'_U, X_U)$. Conversely, the generator function f with parameter γ learns to maximize $l_P(X'_P, X_P)$ and minimize $l_U(X'_U, X_U)$ respectively along with the objective of UAE. In this setting, the neural network parameters $\gamma, \gamma_P, \gamma_U$ all are trained together to solve the following optimization problem shown in Eq.(2.2).

$$\max_{\gamma} \Bigg\{ \mathbb{E}_{Q_\phi(X,Z)} \left[\log p_\theta(x|z) \right] + \lambda_P \min_{\gamma_P} \left\{ \mathbb{E} \left[l_P \left(X'_P, X_P \right) \right] \right\} - \lambda_U \min_{\gamma_U} \left\{ \mathbb{E} \left[l_U \left(X'_U, X_U \right) \right] \right\} \Bigg\}, \tag{2.2}$$

$$\text{where } X'_P = a_p(f(X, X_P, X_U)) \ , \ X'_U = a_u(f(X, X_P, X_U)).$$

The pseudocode for our implementation to address the optimization problem depicted in Equation (2.2) is shown in Algorithm 1.

Algorithm 1 UAE-PUPET – pseudocode for training and testing for data-type-ignorant and data-type-aware conditions

// One step training
Step 1:
$\hat{\mathbf{X}} \leftarrow \mathbf{f}(X, X_P, X_U)$
Step 2:
$X'_P \leftarrow a_p(\hat{\mathbf{X}})$, then calculate cross entropy loss i.e. l_P
$X'_U \leftarrow a_u(\hat{\mathbf{X}})$, then calculate cross entropy loss i.e. l_U
Step 3:
Calculate generator loss and update γ.
// To calculate generator loss or (gen_loss):
calculate *reconstruction loss* i.e. $\text{mse}(X, \hat{\mathbf{X}})$
$gen\_loss \leftarrow$ *reconstruction loss* $+ (\lambda_U * l_U) - (\lambda_P * l_P)$
Back propagate and update γ parameters.

Step 4: Take Step 1, then $X'_P \leftarrow a_p(\hat{\mathbf{X}})$, then calculate private_loss i.e. l_P
Back propagate and update γ_P parameters.
Step 5: Take Step 1, then $X'_U \leftarrow a_u(\hat{\mathbf{X}})$, then calculate utility_loss i.e. l_U
Back propagate and update γ_U parameters.

// Testing
Take **Step 1** to generate privatized data $\hat{\mathbf{X}}$.
if data-type-ignorant condition **then**
    pass // $\hat{\mathbf{X}}$ without data type constraints
else if data-type-aware condition **then**
    for each categorical variable $\hat{\mathbf{X}}[i:j]$ **do**
        $\hat{\mathbf{X}}[i:j] \leftarrow enforce\_constraints(\hat{\mathbf{X}}[i:j])$
        // $enforce\_constraints$ assigns 1 to argmax index
        // and assigns 0 to other indices
    end for
end if
$X'_P \leftarrow a_p(\hat{\mathbf{X}})$, then evaluate private accuracy.
$X'_U \leftarrow a_u(\hat{\mathbf{X}})$, then evaluate utility accuracy.

// Recall privatized data is tested under multiple adversaries.

2.3 Experiments and Results

We perform comprehensive experiments on four widely-used datasets such as MNIST, Fashion MNIST, UCI-adult, and US Census Demographic Data, to demonstrate the effectiveness

of the proposed privacy mechanism i.e. solving the optimization problem in Eq.(2.2). Firstly, we discuss the results for data-type-ignorant conditions and compare them to existing works wherever possible and then present the result of data-type-aware conditions. We perform experiments with hyperparameter $\lambda_P = [0,10,20,...,100]$ for data-type-ignorant conditions and plot a UPT (Utility Privacy Tradeoff) curve. The UPT curve is used to represent the upper bound on the performance of the privacy-utility mechanism. As such, it consists of the upper convex hull of all the achieved operational points. The operational interpretation of the upper convex hull relies on an operational interpretation of any line connecting two operational points. Recall that each operational point is defined by two accuracy levels, achieved by averaging over an entire test dataset. If we split the test dataset in two parts of sizes α and $1-\alpha$ times the original size, respectively, and apply the first part to the mechanism achieving operational point P_1, and the second part to the mechanism achieving P_2, then the average over the entire dataset should achieve operational point $\alpha P_1 + (1-\alpha)P_2$. It is in this sense that the upper convex hull is achievable. Notice that for each λ_P value, we conduct 25 experiments with random weight initialization and then report the mean for that particular λ_P. Similarly, the *best guess point* refers to a certain accuracy that can be achieved by guessing a class which has the highest frequency. The pseudocode for our implementation is shown in Algorithm 1. For more information about the architecture of the generators, discriminators, and different hyperparameters used for multiple experiments, please refer to our source code in the following repository: *https://github.com/bishwasmandal246/uae-pupet*

2.3.1 MNIST

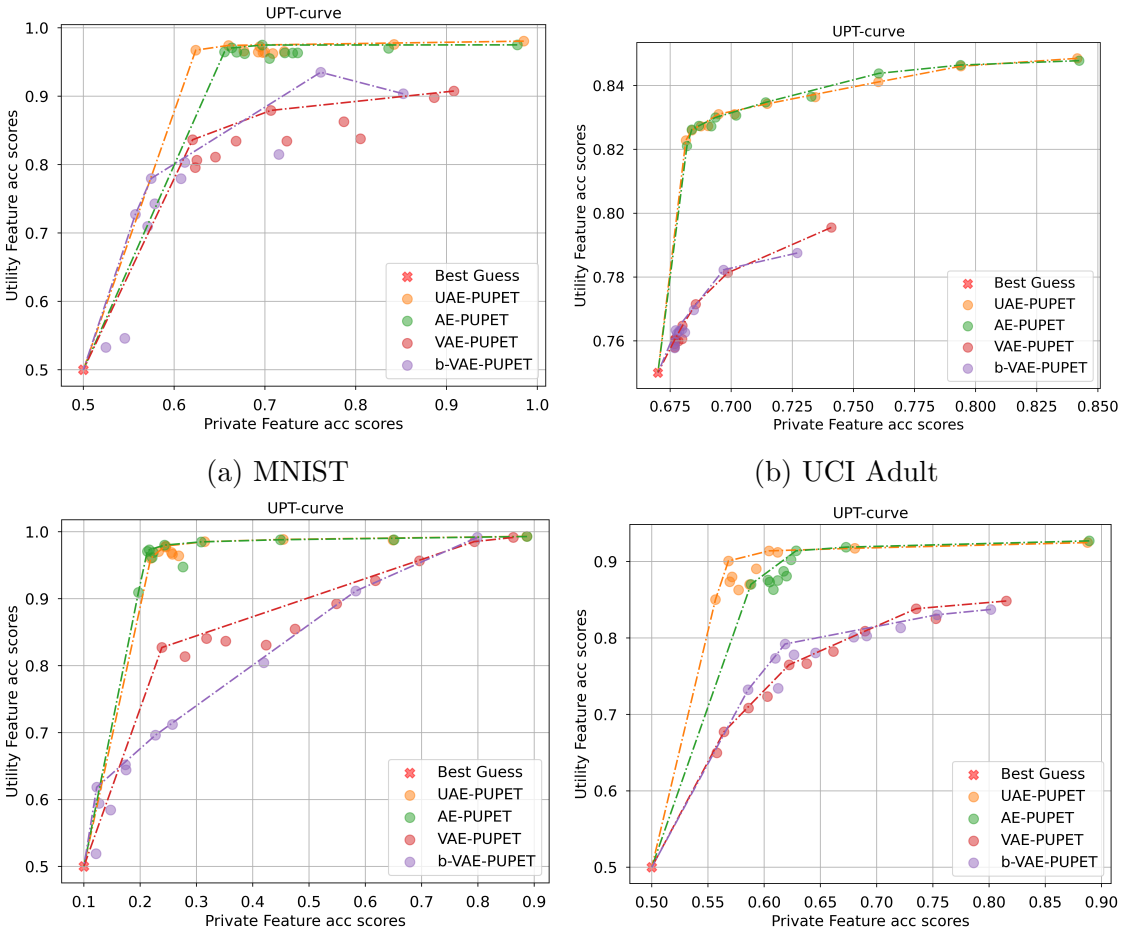
This experiment considers the same setting as Chen et al. (2019) and Rezaei et al. (2018) where the private attribute refers to whether the number is odd or even, while the utility attribute encodes whether the number is ≥ 5 or not. Figure 2.1, shows the original and privatized images generated by the proposed privacy mechanism *UAE-PUPET*. We also implemented our privacy mechanism using autoencoders (AE-PUPET), variational autoen-

Models	Private attr.		Utility attr.	
	accuracy	**AUROC**	**accuracy**	**AUROC**
without distortion(raw)	0.98	0.98	0.98	0.98
emb-g-filter (Chen et al., 2019)	0.651	-	0.855	-
VAE-PUPET	0.620	0.619	0.836	0.834
β-VAE-PUPET	0.6117	0.6119	0.8032	0.81
AE-PUPET	0.654	0.657	0.964	0.964
UAE-PUPET	0.625	0.622	0.967	0.967

Table 2.1: MNIST accuracy and auroc results comparison

coders (VAE-PUPET), and β-VAE (Higgins et al., 2017) (β-VAE-PUPET) to demonstrate empirically that removing the Gaussian restrictions from the latent variables (UAE-PUPET) is helpful in achieving better privacy and utility guarantee i.e. low private attributes accuracy and high utility attributes accuracy. Note that, β-VAE is an extension of VAE with $\beta > 1$ as the multiplier to the KL divergence term of VAE's learning objective. Table 2.1 illustrates the upper bound accuracy, and area-under receiver operating characteristic curve (AUROC) results of our privacy mechanism *UAE-PUPET*, and its comparison to emb-g-filter from Chen et al. (2019), AE-PUPET, VAE-PUPET and β-VAE-PUPET. The results for Chen et al. (2019) are extracted from the paper, where the authors only provide accuracy scores, and hence AUROC is kept blank. Our proposed mechanism clearly outperforms emb-g-filter by achieving 2.6% less (absolute) accuracy on the private attribute and 11.2% higher (absolute) accuracy on the utility attribute even when UAE-PUPET is tested under multiple adversaries (weak adversary, strong adversary, and the adversaries trained during the privacy mechanism training) whereas the emb-g-filter (Chen et al., 2019) is tested only under their original restrictive single adversary model. Similarly, with our setup, it is clear to see the use of UAE also provides better privacy and utility guarantees than that of AEs, VAEs or β-VAEs. Figure 2.3a shows the UPT curve and compares different versions of PUPET. Each point in the curve represents the mean of 25 experiments for a particular λ_P value. Note that the points in the upper-left (North-West) regions are the most favorable.

In Figure 2.4, we provide a visualisation of latent variables before and after the applica-



(a) MNIST

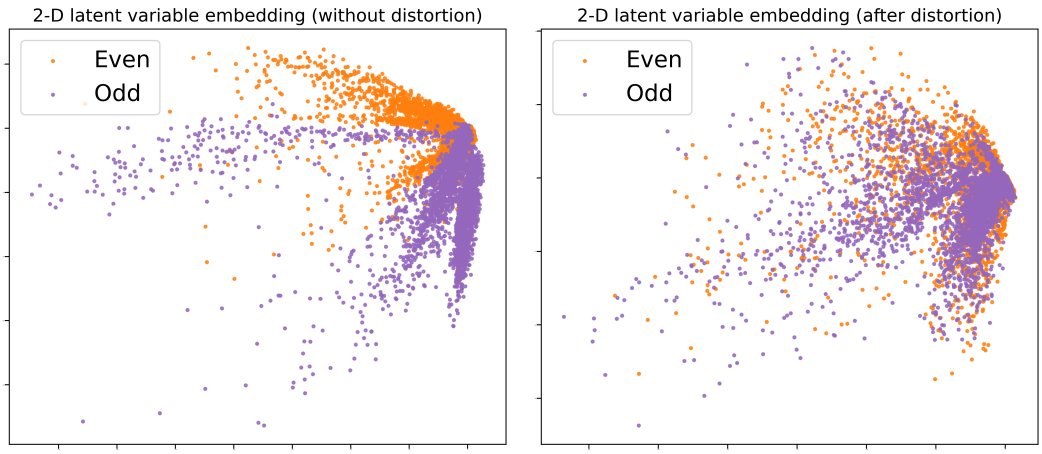
(b) UCI Adult

(c) Fashion MNIST

(d) US Census

Figure 2.3: UPT Curves: Utility Privacy Tradeoff curves

tion of our privacy mechanism when the dimension of the latent variable is two. In Figure 2.4, we can observe distinct regions in the two-dimensional space for the private feature – even or odd, which makes the correct prediction of private feature possible for the adversary. However, after applying the privacy mechanism, it can be seen that the private feature – even or odd – doesn't have distinct regions in the two-dimensional space and the data points under a similar region sometimes correspond to even and sometimes to odd, making the correct prediction of the private feature difficult even for the *best performing adversary*– an adversary that classifies the private feature with the highest accuracy.



(a) before UAE-PUPET (b) after UAE-PUPET

Figure 2.4: MNIST: Latent geometry visualisation

2.3.2 UCI Adult

Models	Private attr.		Utility attr.	
	accuracy	**AUROC**	**accuracy**	**AUROC**
without distortion(raw)	0.83	0.75	0.84	0.76
VFAE (Louizos et al., 2017)	0.802	0.703	0.851	0.761
LMFIR (Song et al., 2018)	0.728	0.659	0.829	0.741
emb-g-filter (Chen et al., 2019)	0.717	0.632	0.822	0.731
VAE-PUPET	0.698	0.573	0.781	0.597
β-VAE-PUPET	0.696	0.58	0.782	0.607
AE-PUPET	0.6819	0.525	0.821	0.71
UAE-PUPET	0.6814	0.521	0.822	0.72

Table 2.2: UCI Adult accuracy and AUROC result comparison

In this experiment, we set the private feature as gender, and utility feature as income. Data pre-processing steps include removing all the data points with missing values, converting categorical variables to one-hot encoding representations and normalizing all the variables. We compare the proposed privacy mechanism results to Variational Fair AutoEncoder (VFAE) (Louizos et al., 2017), Lagrangian Mutual Information-based Fair Representations (LMIFR) (Song et al., 2018) and emb-g-filter (Chen et al., 2019). Similar to the MNIST experiment we also implement AE-PUPET, VAE-PUPET and β-VAE-PUPET. It

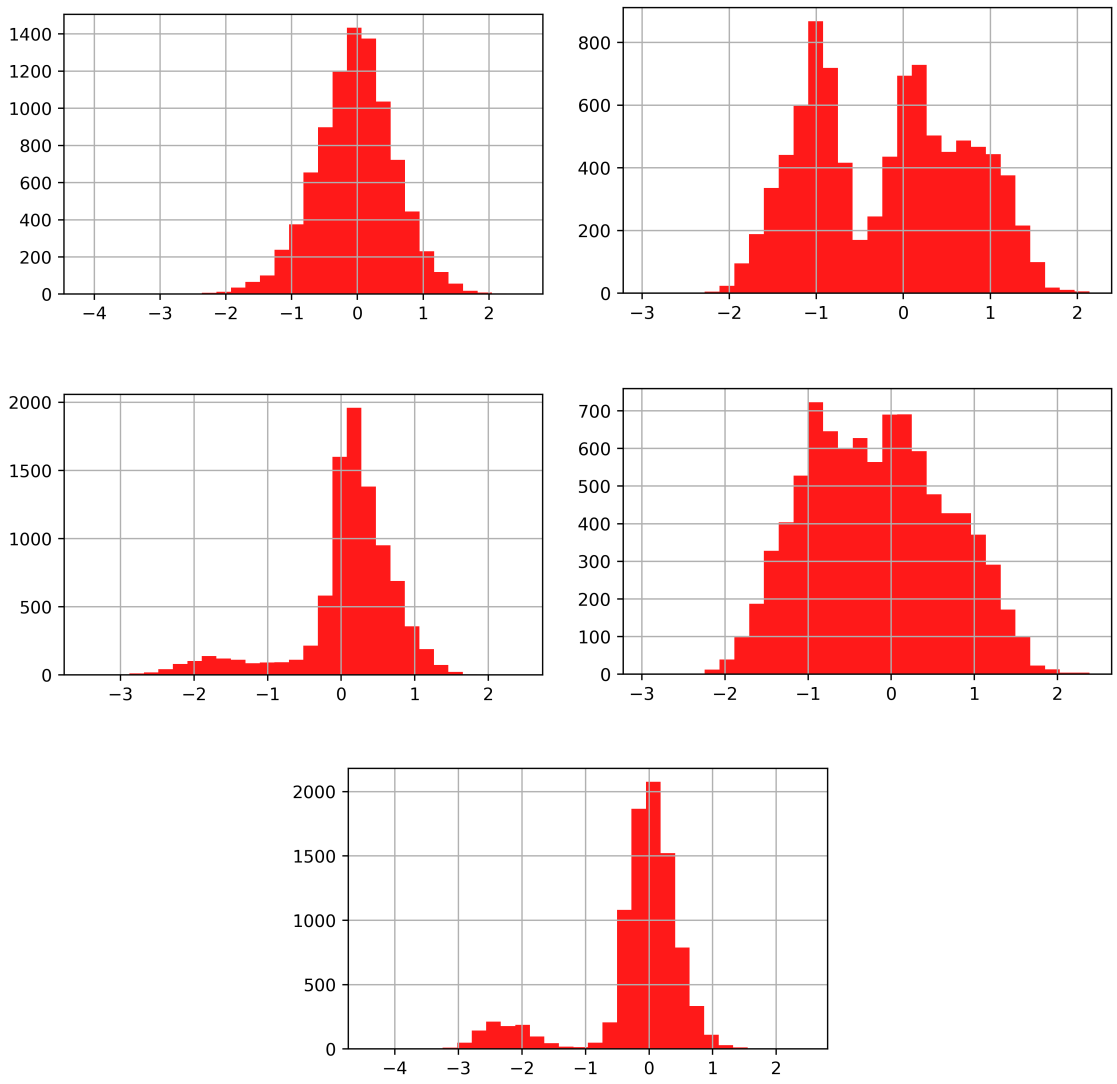


Figure 2.5: UCI Adult: Different distributions of distortions after using UAE-PUPET

is evident through Table 2.2 that our proposed method UAE-PUPET has the least accuracy score of 0.6814, and AUROC score of 0.521 for the private feature and comparable accuracy score of 0.822, and AUROC score of 0.72 for the utility feature even after being tested under multiple adversaries. Figure 2.3b shows that UAE-PUPET and AE-PUPET perform quite similarly for this dataset.

Recall that our privacy mechanism does not add external noise during training. However, the regularization term provided by the discriminator losses when updating the γ parameters

reflect a distortion. Figure 2.5 illustrates the most common distortion distributions observed on the UCI Adult dataset when privatizing using UAE-PUPET. By not forcing any specific noise distribution when training, the privacy mechanism has a higher degree of freedom to produce different distributions of distortion as shown in Figure 2.5 to ensure better privacy and utility guarantees.

2.3.3 Additional Results

Figure 2.6: **Fashion MNIST**: Odd and even adjacent columns show original and privatized versions (UAE-PUPET) respectively.

Besides MNIST and UCI Adult, we also conduct experiments on Fashion MNIST and US Census Demographic Data. These experiments employ private and utility features that do not compare to any existing method; they are being presented solely for the purpose of demonstrating how different PUPET models worsen the adversaries' results. For Fashion MNIST dataset, the private feature is the identity of the fashion article (T-shirt/top, Trouser, Pullover, Dress, Coat, etc.) and the utility feature is encoded on two labels: Upper (meaning T-shirt/top, Pullover, Dress, Coat, Shirt) and Miscellaneous. Figure 2.6, shows the privatized images, along with their original versions. We see that the privatized images appear to have been changed to different articles of clothing. We also notice blurriness of privatized images, sometimes appearing to be comprised of two different images juxtaposed on one. Experiment in this work shows a drop of inference accuracy on private feature from

98% to 21% using UAE-PUPET and AE-PUPET, 23% using VAE-PUPET, 41% using β-VAE-PUPET under *best performing adversary* whereas the inference accuracy on the utility feature decreases slightly from 99% to 96% using UAE-PUPET, 97% using AE-PUPET, 82% using VAE-PUPET and 80% using β-VAE-PUPET.

Similarly, US Census Demographic Data is the American Community Survey data for 2017 that consists of 74,001 records for different counties. It has a total of 37 features, out of which we select the sixteen features which are highly correlated to each other, similar to the setting in Sharma et al. (2021). Examples of some of the selected features include the men population, women population, population of citizens eligible to vote, per capita income, percentage of population unemployed etc. Among the sixteen features, we select Employed as the utility and Income as the private feature. Unlike Sharma et al. (2021), we further categorize the utility feature into two labels i.e. ≤ 2000 or > 2000 and private feature into two labels i.e. ≤ 55000 or > 55000 to make it a balanced classification problem. All the other fourteen features are continuous, and thus we normalize them based on their mean and standard deviation. Similarly, we ignore data points that have missing values. We use a total of 43,657 data points for training, and 29,105 data points for testing purposes. Our experiments show that the accuracy for private features drops down from 88% to 56% using UAE-PUPET, 58% using AE-PUPET, 68% using VAE-PUPET, 61% using β-VAE-PUPET, and the utility accuracy drops from 92% to 90% using UAE-PUPET, 87% using AE-PUPET and 80% using VAE-PUPET, and 79% using β-VAE-PUPET. This experiment again clearly shows better performance of UAE-PUPET over other models. Please refer to Figure 2.3c and Figure 2.3d for the complete comparison of these PUPET models.

2.3.4 Data-Type-Aware Condition

In this section, we discuss the data-type-aware condition and its result. We concentrate on the UCI Adult dataset since it is the only one out of four datasets used, that contains categorical variables. The exponential mechanism (McSherry and Talwar, 2007) in differ-

Models	Private attr.		Utility attr.	
	accuracy	**AUROC**	**accuracy**	**AUROC**
without distortion(raw)	0.83	0.75	0.84	0.76
VAE-PUPET	0.6843	0.5806	0.7831	0.6272
β-VAE-PUPET	0.6791	0.53	0.771	0.5813
AE-PUPET	0.6832	0.5414	0.8245	0.728
UAE-PUPET	0.6817	0.5361	0.8220	0.7151

Table 2.3: UCI Adult: data-type-aware conditions results

ential privacy is designed specifically for discrete/categorical variables where classes are selected based on a scoring function. Similarly, Giggins and Brankovic (2012) proposed a noise addition technique to the categorical data. However, these formulations solve different problem than ours. To the best of our knowledge, the existing literature on using an adversarial learning framework to solve the privacy-utility tradeoff problem does not discuss data-type-aware conditions which is to say, correctly encoding the categorical feature as one hot encoding that represents exactly one class, e.g.(a) (say [0, 0, 0, 1, 0]) after the distortion of data through the privacy mechanism. Instead, the privacy mechanism outputs different probability values for each class, e.g. (b) (say [0.1, 0.05, 0.03, 0.8, 0.02]) and this difference between (a) and (b) is referred to as noise (Edwards and Storkey, 2016; Madras et al., 2018; Chen et al., 2019). However, this notion of addition of noise for categorical features is heavily flawed as the user cannot specify a categorical variable as 10% of the first class, 5% of the second class, and so on. Therefore, we formulate data-type-aware conditions which use the precise representation of categorical variables before disclosing the data (refer to Algorithm 1).

Finally, we discuss about the experimental results for data-type-aware conditions on different versions of PUPET. Table 2.3 shows that, even when enforcing constraints on the categorical variables, our PUPET models are capable of providing similar privacy and utility guarantees compared to the data-type-ignorant conditions. Additionally, we notice that UAE-PUPET and AE-PUPET results are again very similar in performance. With more

experiments we observed that having small variance improves the performance slightly for this dataset which explains the similarity in results for UAE and AE based PUPETs – recall that with zero variance, UAE's learning objective is same as AE's.

2.4 Conclusion

In this chapter, we introduced a novel UAE-based privacy mechanism (UAE-PUPET), and showed that it can attain better privacy-utility tradeoffs than the existing works. Additionally, the results for AE-PUPET were found to be very close to UAE-PUPET, but the use of VAE and β-VAE showed poor results on all the datasets. UAE-PUPET is still better than AE-PUPET in the sense that it can behave like an AE-PUPET if required but the converse is not true. Better overall results with UAE-PUPET and AE-PUPET imply that forcing a Gaussian distribution on the latent variable of autoencoders (such as in VAE and β-VAE) appears to hinder the privacy mechanism. Our privacy mechanism is focused on end-to-end stochastic mapping of input to the output. However, VAE and β-VAE have a KL-divergence term in their learning objective, which deviates from only focusing on the reconstruction term. The results for VAE and β-VAE on UCI Adult is particularly poor because of the heterogeneous nature of the data. Ma et al. (2020) formulate VAEs to handle the heterogeneous nature of data but this is a two stage process which adds complexity and is harder to incorporate with our settings. Instead, the use of UAE helps deal with heterogeneous data types, behaves like an AE if required, and can still add randomness in the latent space like VAE and β-VAE.

Chapter 3

Optimizing Privacy and Utility Tradeoffs for Group Interests Through Harmonization

3.1 Introduction

In numerous practical applications, tabular data is routinely generated as a standard output of relational database systems (Shwartz-Ziv and Armon, 2021). This chapter primarily addresses scenarios wherein data analysts undertake the meticulous examination of tabular datasets to formulate predictive models for diverse classification endeavors in a two-group setting. The process of constructing such predictive models entails two principal methodologies: analysts may opt to manually annotate the data or alternatively pursue and amass an auxiliary dataset possessing identical input features and encompassing the labels targeted for prediction. Beyond the predictive utility of these models, analysts possess the capability to develop models for inferring additional private or sensitive features through analogous methodologies. Hence, it becomes apparent that while machine learning models serve constructive purposes, they also harbor the potential for malicious exploitation in extracting

sensitive or private information.

To mitigate the associated privacy risk, existing literature suggests various privacy mechanisms such as Edwards and Storkey (2016); Huang et al. (2017); Madras et al. (2018); Pittaluga et al. (2018); Osia et al. (2020); Huang et al. (2019); Chen et al. (2019); Wu et al. (2020); Morales et al. (2020); Xiao et al. (2020); Erdemir et al. (2021); Wu et al. (2021); Mandal et al. (2022) designed to protect sensitive information while still enabling the prediction of useful feature. These mechanisms frequently use adversarial optimization techniques to train their generative models for the data sanitization process. The adversarial optimization methods can be trained either using the auxiliary dataset, if available, or by initially annotating the private and utility labels for the given input data and then training the privacy mechanism. The overarching goal is to develop a privacy mechanism that minimizes the mutual information between the input data and sensitive features while maximizing the mutual information with utility features to safeguard privacy and still enable accurate inferences of utility features.

Previous studies have emphasized privacy (*inference privacy*) concerns associated with analyzing datasets that assume all users share the same private and utility characteristics. However, in real-world situations, it is unrealistic to expect uniform privacy requirements among all individuals within a dataset (Zhang and Zhao, 2007). Consequently, it is likely that diverse user populations, such as those on the internet, possess varying private and utility attributes. These distinct user groups, each with their unique sets of private and utility attributes, are closely interconnected, enabling the transfer of information from one group to another. For instance, collaborative recommendation systems may suggest new items to one group of customers based on the purchase history of another group of customers (Adomavicius and Tuzhilin, 2005). While this collaborative use of data can have beneficial outcomes, it can also inadvertently reveal sensitive information. For instance, Irani et al. (2011) demonstrated that utilizing information from an individual's social network connections can effectively disclose details like their location, sexual orientation, marital status,

and political affiliation.

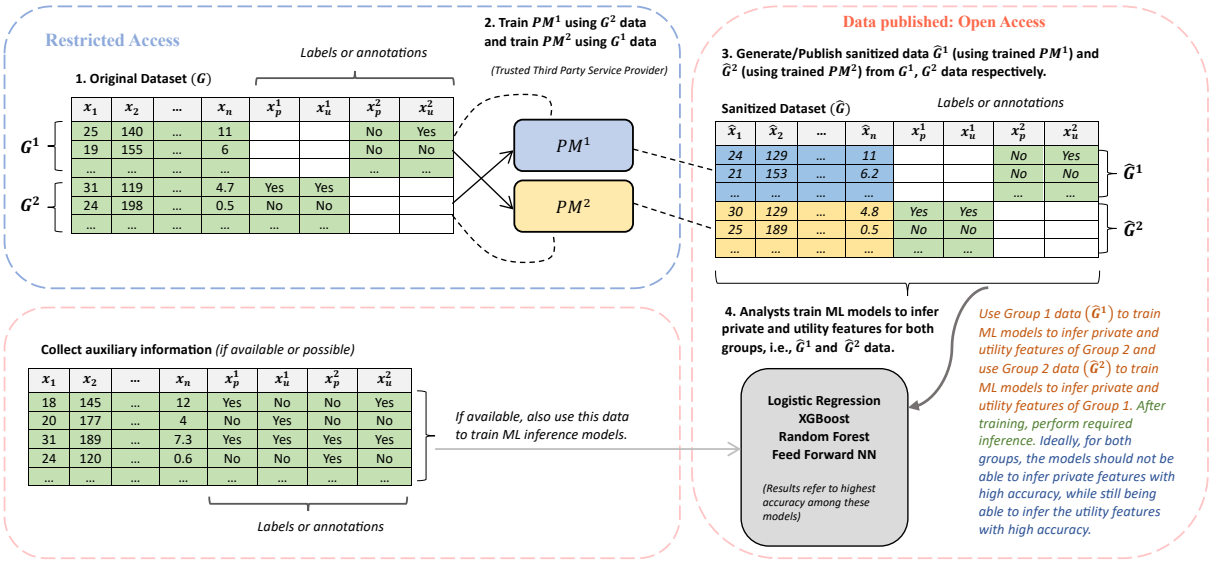


Figure 3.1: Overview of the privacy-utility tradeoff in two-user group setting. *Restricted Access* block demonstrates how data from one group trains the privacy mechanism for another and vice-versa. *Open Access* block details the public release of sanitized data available for public analysis. Blank spaces represent private and utility features of a particular group which are not present in the dataset, and analysts or adversaries aim to make correct predictions of these features.

In this work, we introduce a novel problem formulation that revolves around a tabular dataset comprising two distinct user groups, each characterized by unique private and utility features. The primary objective is to achieve highly accurate predictions of the utility features for both user groups while simultaneously ensuring that accurate predictions of their private features are highly improbable. While it is acknowledged that there will be cases where users' private attributes could be correctly inferred, our aim is to keep the accuracy of such predictions intentionally low, thus granting users plausible deniability. In contrast to prior approaches, where tabular datasets typically featured a single set of private and utility features, and the privacy mechanisms being trained relied on auxiliary datasets or manual annotation of data points, our approach has the flexibility to accommodate multiple user groups with varying private and utility features. In this study, for simplicity, our attention

is directed towards a scenario characterized by the involvement of two distinct user groups. Importantly, our privacy mechanism's training does not depend on manual annotation or auxiliary datasets; it solely relies on the data originating from the two user groups.

3.2 Related Work

In the realm of privacy, there exist two primary categories (Sun et al., 2018; Sun and Tay, 2020). The first category, known as *data privacy*, is primarily focused on safeguarding the original, unprocessed data. In contrast, *inference privacy* seeks to shield sensitive information that can be deduced or inferred from the disclosed raw data, often as a result of correlations. *Raw data*, in this context, refers to the initial, unaltered data collected or generated, representing data in its fundamental state prior to any form of examination, modification, or inference.

Classical methods to preserve data privacy encompass Differential Privacy (DP) (Dwork and Roth, 2014), and Homomorphic Encryption (HE) (Acar et al., 2018). DP primarily ensures that databases with a single differing record remain indistinguishable, primarily serving as a defense against membership inference attacks. On the other hand, HE allows computations to be performed on encrypted data. It is important to note that these methods do not directly address the issue of inference privacy, which constitutes the central focus of our research. Label Differential Privacy (Wu et al., 2022) introduces noise to labels to safeguard them, but it does not apply noise to raw data to prevent the inadvertent disclosure of private labels, as is the specific concern within inference privacy.

In a broader context, while differential privacy (DP) doesn't inherently target the challenge of inference privacy, specific adjustments can enable DP to provide a formal guarantee of inference privacy under certain assumptions (Ghosh and Kleinberg, 2017). However, determining the optimal level of noise to introduce into higher-dimensional data for achieving differential privacy can be impractical. Homomorphic Encryption (HE) can be used

for inference privacy with certain assumptions, but conducting computations on encrypted data becomes intricate, especially with complex neural network calculations (Chen et al., 2022). Therefore, recent research on inference privacy has utilized adversarial optimization techniques to minimize the mutual information between disclosed data and private features in various contexts (Edwards and Storkey, 2016; Huang et al., 2017; Madras et al., 2018; Pittaluga et al., 2018; Osia et al., 2020; Huang et al., 2019; Chen et al., 2019; Wu et al., 2020; Morales et al., 2020; Xiao et al., 2020; Erdemir et al., 2021; Wu et al., 2021; Mandal et al., 2022). These studies have primarily focused on a single user group i.e., all users with identical private and utility attributes.

Within the field of data privacy, substantial investigation has been undertaken regarding strategies pertaining to multi-group or multi-user privacy. Various methods involving secure multi-party computation (SMPC) (Knott et al., 2022) are utilized to preserve data privacy in scenarios featuring multiple parties, ensuring that individual parties are unable to access the data of others. Heterogeneous Differential Privacy (Alaggan et al., 2015) takes into account the varying privacy expectations of different users, while Personal Differential Privacy (Ebadi et al., 2015) allows for individualized privacy budgets for each user, albeit with a trade-off of reduced data utility. To address this challenge, Niu et al. (2021) introduced a utility-aware sampling method that caters to the diverse privacy requirements of different users. Another approach, as outlined in Koufogiannis and Pappas (2016), presents a model wherein multiple data owners with private real-valued data are shielded using a Gaussian mechanism against multiple data users seeking to glean answers to linear queries. Furthermore, Koufogiannis and Pappas (2018) provides a differentially private response that ensures the extent of private data leakage depends on the proximity of two users within the network. A comprehensive multi-user privacy approach is presented in Ghazi et al. (2022), which combines counting Bloom filters (Bloom, 1970), Differential Privacy (DP), homomorphic encryption (HE), and secure multi-party computation (SMPC) (Evans et al., 2018) to address cardinality and frequency histogram problems.

Despite the considerable body of work dedicated to the exploration of multi-user privacy within the context of data privacy, it is worth noting that, to the best of our knowledge, there exists a noticeable gap in research concerning multi-user inference privacy, particularly when dealing with distinct sets of private and utility features. An investigation presented in Wang et al. (2021) delves into privacy-utility tradeoffs within multi-agent systems, wherein each agent employs linear compression and noise perturbation techniques to safeguard private information while still facilitating the accurate identification of utility features. In this system, multiple agents operate independently, assuming a scenario devoid of a central trusted entity. However, it's important to distinguish this concept of *multi-agent* from our setup, as in their case, all agents possess uniform private and utility features, whereas we specifically consider varying private and utility features across different user groups. In the subsequent section, we furnish a comprehensive definition of our privacy-utility tradeoff problem involving two user groups.

3.3 Problem Formulation

Consider a scenario where a dataset G comprises of n features, and is denoted as $\mathcal{X} = \{x_1, x_2, x_3, \cdots, x_n\}$. In addition, the dataset G includes labels or annotations represented as $\{x_p^1, x_u^1, x_p^2, x_u^2\}$, where $x_p^1 \notin \mathcal{X}, x_u^1 \notin \mathcal{X}, x_p^2 \notin \mathcal{X}, x_u^2 \notin \mathcal{X}$. These features or labels may encompass demographic variables such as gender and race, socio-economic metrics like education level and earnings, or digital interactions like mouse movements, link clicks, and content preferences, depending on the specific application domain. Within this dataset, we assume two distinct user groups: Group 1 (G1) and Group 2 (G2), each consisting of $k/2$ rows, resulting in a total of k rows in the entire dataset G. Specifically, G1 includes features and labels denoted as $\mathcal{A} = \mathcal{X} \cup \{x_p^2, x_u^2\}$ and does not contain $\{x_p^1, x_u^1\}$. Here, x_p^1 is G1's private feature, intentionally withheld from disclosure, while x_u^1 represents the label that G1 aims to predict with high accuracy. Similarly, G2 comprises features and labels

denoted as $\mathcal{B} = \mathcal{X} \cup \{x_p^1, x_u^1\}$ and does not include $\{x_p^2, x_u^2\}$. In this context, x_p^2 serves as G2's private feature, deliberately omitted from the dataset, while x_u^2 signifies the label that G2 seeks to predict accurately. It is important to note that these private and utility features correspond to certain classification labels. For simplicity, we assume binary classification labels in this work. The arrangement of these features is visually depicted in Figure 3.1, within the *Restricted Access* block.

If the dataset G were to remain unaltered and be publicly accessible, analysts would possess the capability not only to deduce the valuable utility features (x_u^1 for Group 1 and x_u^2 for Group 2) but also to discern the sensitive or private features (x_p^1 for Group 1 and x_p^2 for Group 2). The attainment of this inference is feasible through the application of machine learning (ML) models. As an illustration, to distinguish the private and utility features of G1, ML models can be trained using G2 data, as G2 data contains private and utility labels specific to G1. Similarly, it is possible to achieve a high-precision prediction of the private and utility features of G2. Therefore, the critical necessity lies in the development of a privacy mechanism that can sanitize the dataset such that it allows accurate inference of utility features for both user groups while constraining the accurate inference of private attributes. While recognizing that there may be instances where some users' private attributes can be correctly inferred, our primary aim is to deliberately limit the accuracy of such predictions to provide users with a degree of plausible deniability. In the context of this study, *privacy refers to the adversary's inability to construct a classifier that can predict the private features with a high accuracy, whereas utility is defined as the ability to construct a classifier that can predict utility features with high accuracy.*

3.3.1 Assumptions and Threat Model

In our problem formulation, we posit a fundamental absence of trust between Group 1 and Group 2, resulting in a reluctance to engage in direct data sharing for the training of their respective privacy mechanisms (the details of which will be discussed in Section 3.4.2).

Instead, we introduce the premise of a trusted third-party service provider, which enjoys the trust of both user groups. Consequently, both Group 1 and Group 2 submit their raw data to this third-party service provider. This arrangement alleviates the need for direct data exchange between Group 1 and Group 2 while enabling the third-party to utilize the data for privacy mechanism training. It is important to emphasize that within our assumptions, the third-party service provider neither has access to any raw auxiliary dataset, nor has options to manually annotate certain portion of the data first. This assumption can be attributed to various reasons, including the absence of such a dataset, its non-public status, or the substantial cost and time required for its acquisition or annotation. Therefore, the provider is restricted to utilizing the data provided by the two user groups exclusively. Additionally, we assume that this third-party provider abstains from providing the user groups with trained privacy mechanism models. Instead, the third-party service provider exclusively provides black-box access i.e. it only delivers the resulting sanitized (privatized) output data produced by the privacy mechanism model when given raw input data. This precautionary measure potentially mitigates extraction of data that was used to train the model. Hence, our framework operates akin to the principles observed in Machine Learning as a Service (MLaaS) applications, where users provide raw data and receive predictions, in our case, both user groups contribute raw data and receive sanitized data.

The raw data remains in a *Restricted Access* block, accessible solely to a trusted third-party service provider until it undergoes sanitization (privatization). It is only when the privacy mechanism is trained and the raw data is sanitized that the data becomes available to the public or analysts. The sanitized dataset, denoted as $\hat{G}$, is depicted in Figure 3.1 in the *Open Access* block. Therefore, for the training of models aimed at inferring private and utility features for Group 1, analysts can use the sanitized G2 data, referred to as $\hat{G}^2$. Likewise, analysts can infer Group 2's private and utility features by training models from $\hat{G}^1$ data.

3.3.2 Philosophical Disagreements

From a philosophical perspective, a realm of discourse emerges. Some individuals might assert that for specific groups of users, only the private features should differ while keeping the utility features constant, while others may propose the opposite, suggesting that only the utility features should vary while maintaining a consistent private feature. This choice is deeply rooted in philosophical contemplations and remains devoid of a universally definitive resolution. In our specific case, we assume the distinctiveness of all private and utility features.

3.3.3 Real World Application

Our problem formulation has myriad potential real-world applications. Consider a scenario in which two hospitals possess datasets containing distinct sets of features. The first hospital seeks to predict a valuable feature that the second hospital possesses, while conversely, the second hospital aims to predict a different valuable feature possessed by the first hospital. Recognizing the costliness of data annotation or finding auxiliary datasets, both hospitals opt to mutually share their data. However, they soon realize that each hospital also hold labels for a feature that is private to them. Without privacy mechanisms, the risk arises that one hospital can infer the private features of the other's data during this exchange. Therefore, they opt for a collaborative approach where they refrain from direct data sharing and instead entrust their data to a trusted third-party service provider for sanitization and subsequent publication. Following data publication, both hospitals can leverage each other's data for predictive purposes without requiring additional data annotation, all while safeguarding the privacy of their respective private features. Likewise, there are numerous application domains where valuable inference privacy challenges can be effectively addressed by our setup.

3.4 Methodology

In this section, we present our proposed methodology for sanitizing the dataset for two-group scenario. The primary objective of this methodology is to enable the precise prediction of utility features for both user groups while concurrently maintaining the privacy of these distinct groups. Our approach involves the introduction of a data-sharing mechanism that is capable of accommodating existing privacy mechanisms. First, we provide an overview of privacy mechanism ALFR (Edwards and Storkey, 2016) from existing literature and UAE-PUPET (Mandal et al., 2022) which was also introduced in the previous chapter, explaining their use in data sanitization for a single-group scenario. Later, we adapt these mechanisms for a two-group scenario.

3.4.1 ALFR and UAE-PUPET

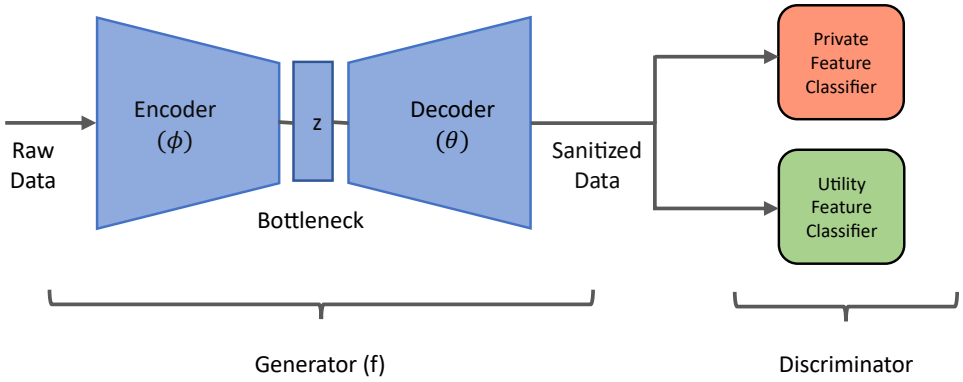


Figure 3.2: ALFR and UAE-PUPET architecture

ALFR (Adversarial Learned Fair Representations) (Edwards and Storkey, 2016) and UAE-PUPET (Uncertainty Autoencoder based Privacy and Utility Preserving end-to-end Transformation) (Mandal et al., 2022) are neural network architectures composed of two main components: a generator and a discriminator (refer to Figure 3.2). The generator

comprises a generative model consisting of an encoder-decoder pair, while the discriminator includes classifiers for predicting private and utility features. Both ALFR and UAE-PUPET were designed to address privacy-utility tradeoffs in single-group scenarios.

In the context of a single group, it is assumed that a third-party service provider, responsible for training the privacy mechanism, has access to a raw auxiliary dataset (belonging to a virtual group $G3$) denoted as G^3. The privacy mechanism (PM) is trained using data from the G3 dataset, and data from G1 is sanitized after the training of PM^1, where $G1$ represents the sole group in this single-group scenario, and PM^1 represents PM responsible for sanitizing G1 data.

The generator, in its most general form, is represented as a function f, and its weights are denoted as $\gamma = (\phi, \theta)$. It takes the training input as $G^3_{\mathcal{X}}$, where the superscript denotes the group and the subscript denotes the feature(s), and generates sanitized training data $\hat{G}^3_{\mathcal{X}}$ as follows: $\hat{G}^3_{\mathcal{X}} = f(G^3_{\mathcal{X}})$. For the generator part, ALFR leverages a standard autoencoder (AE), while UAE-PUPET uses an Uncertainty Autoencoder (UAE) (Grover and Ermon, 2018). UAE introduces randomness and does not make any assumptions about Gaussian or any prior distribution on the latent variable or bottleneck, unlike the Variational Autoencoder (Kingma and Welling, 2022). The degree of randomness can be adjusted using a hyperparameter specified by the user. Both ALFR and UAE-PUPET utilize neural networks, denoted as functions s_p and s_u, which act as simulated adversaries and utility provider networks and predict private and utility features represented as $\hat{G}^3_{x^1_p}, \hat{G}^3_{x^1_u}$, respectively. The weights of s_p, s_u are represented as γ_p and γ_u, respectively. The simulated adversaries and utility provider's network contribute essential loss during the training stage. The scalars α, λ_p, and λ_u are hyperparameters that determine the weighting of the competing objectives. This inclusion serves the purpose of introducing implicit regularization, involving noise introduction. This noise is intended to make the classification of private features challenging while allowing the classification of utility features. It is important to note that the introduced noise required for the task is solely influenced by the losses contributed

by these discriminator models, and no additional noise is introduced. The optimizations within Algorithm 2 provide a comprehensive view of the operational intricacies of ALFR and UAE-PUPET for the sake of completeness. In the upcoming section, we present our proposed *data sharing mechanism* that extends the existing privacy techniques ALFR, and UAE-PUPET for managing privacy-utility tradeoffs in a two-group setting.

Algorithm 2 ALFR and UAE-PUPET training algorithm

$\gamma, \gamma_p, \gamma_u \leftarrow$ initialize weights
$U \leftarrow True$
repeat
  $\hat{G}^3_{\mathcal{X}} \leftarrow f(G^3_{\mathcal{X}})$, $\hat{G}^3_{x^1_p} \leftarrow s_p(\hat{G}^3_{\mathcal{X}})$, $\hat{G}^3_{x^1_u} \leftarrow s_u(\hat{G}^3_{\mathcal{X}})$
  $C \leftarrow \text{MSE}(\hat{G}^3_{\mathcal{X}}, G^3_{\mathcal{X}})$
  $l_p \leftarrow \text{cross\_entropy}(\hat{G}^3_{x^1_p}, G^3_{x^1_p})$
  $l_u \leftarrow \text{cross\_entropy}(\hat{G}^3_{x^1_u}, G^3_{x^1_u})$
  $L \leftarrow \alpha C - \lambda_p l_p + \lambda_u l_u$
  if ALFR **then**
    if U **then**
      Update γ_p to *maximize* the loss L
    else
      Update (γ, γ_u) to *minimize* the loss L
    end if
  else if UAE-PUPET **then**
    if U **then**
      Update γ to *minimize* the loss L
    else
      Update γ_p to *minimize* the loss l_p
      Update γ_u to *minimize* the loss l_u
    end if
  end if
  $U \leftarrow$ not U
until Deadline

3.4.2 Data Sharing Mechanism

We can employ a unified adversarial architecture for training the privacy mechanism in a manner that facilitates data sanitization for both user groups. The unified architecture en-

tails the integration of two additional classifiers into the discriminator for the second group's private and utility feature, coupled with the practice of alternating gradient updates between iterations by utilizing data from the alternate group. However, it is worth noting that adversarial privacy mechanisms are known to exhibit stability issues, particularly when the loss function lacks strong concavity (Xing et al., 2021). Utilizing a single adversarial architecture while introducing additional optimizations for the second group may exacerbate these challenges. To confirm, we conducted experiments employing a unified architecture, and indeed, we observed the presence of stability issues. Therefore, we train separate adversarial architecture for each group, denoted as PM^1 and PM^2. In this setup, PM^1 is dedicated to the privatization of G1 data, while PM^2 is responsible for the privatization of G2 data.

The proposed data sharing mechanism follows an iterative and sequential process. Initially, the training of $(PM^1)_m$ for Group 1 is carried out, utilizing data from Group 2, specifically $G^2_{\mathcal{X}}$, along with labels $G^2_{x^1_p}$ and $G^2_{x^1_u}$. The subscript m signifies the iterative process at time m, with $m = 1$ as the starting point. During this training, the generator function $(f^1)_m$ is trained with $(s^1_p)_m$ and $(s^1_u)_m$. It is important to note that $(f^1)_m$ aims to learn how to sanitize data so that only the utility labels can be accurately inferred from the sanitized data. Once the privacy mechanism is trained, $(f^1)_m$ is utilitzed to generate sanitized data for Group 1, denoted as $(\hat{G^1_{\mathcal{X}}})_m$.

Considering that Group 1's data is now sanitized and intended for use by analysts in training predictive classifiers (if the sanitized data were to be published at this time), $(PM^2)_m$ chooses to leverage this sanitized data instead of the original Group 1 data for training its privacy mechanism catering to Group 2. Consequently, the generator function for PM^2_m, denoted as $(f^2)_m$, undergoes training using $(\hat{G^1_{\mathcal{X}}})_m$ data, in conjunction with labels $G^1_{x^2_p}$ and $G^1_{x^2_u}$, and utilizes $(s^2_p)_m$ and $(s^2_u)_m$ for $(PM^2)_m$.

After the training of the privacy mechanism, the generator function $(f^2)_m$ generates sanitized data for Group 2, referred to as $(\hat{G^2_{\mathcal{X}}})_m$. Subsequently, the third-party service

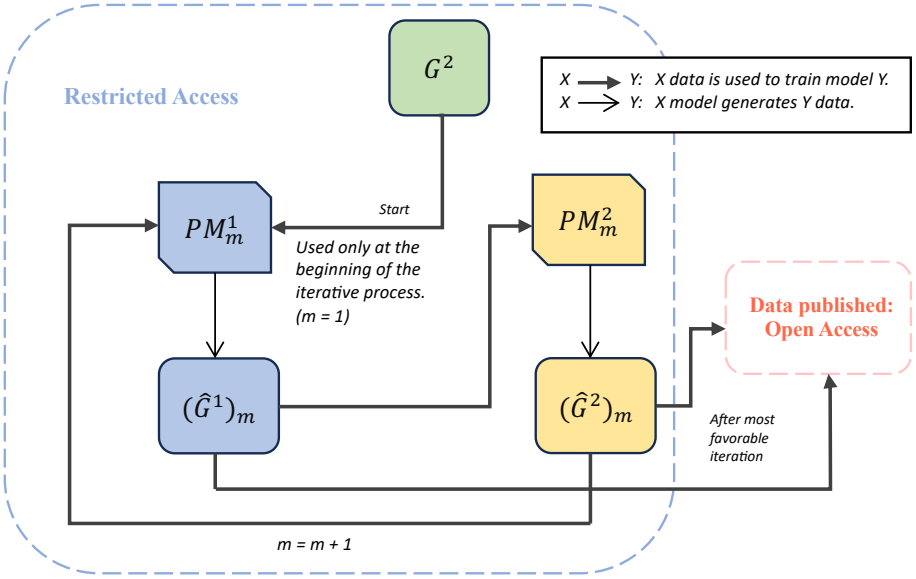


Figure 3.3: Sequential and iterative nature of the data sharing approach.

provider, aware of the alterations in the data used to train the privacy mechanism $(PM^1)_m$, proceeds to retrain this privacy mechanism using the freshly sanitized data from Group 2. For Group 1, the third-party service provider now utilizes the updated sanitized data to retrain the privacy mechanism, denoted as $(PM^1)_{m+1}$, and Group 2 follows suit in retraining its privacy mechanism, referred to as $(PM^2)_{m+1}$, using the updated sanitized data from Group 1. In this context, the subscript $m+1$ signifies another round of the iterative process. This iterative process can be repeated for a total of T rounds, where the index $m = \{1 \cdots T\}$.

The weights associated with the generators are meticulously preserved at each iteration m. The iteration that successfully achieves the desired privacy and utility guarantees is strategically utilized to generate the ultimate sanitized data for both groups, which is subsequently disclosed to the public. The final disclosed data from Group 1 is represented as $\hat{G}^1_{\mathcal{X}} \cup \{G^1_{x_p^2}, G^1_{x_u^2}\}$, while the ultimate disclosed data from Group 2 is denoted as $\hat{G}^2_{\mathcal{X}} \cup \{G^2_{x_p^1}, G^2_{x_u^1}\}$. Note that, the iterative process notation is removed once the process is completed. For a visual representation of the iterative process, please refer to Figure 3.3.

Algorithm 3 Data Sharing Mechanism

$(\gamma^i)_m, (\gamma^i_p)_m, (\gamma^i_u)_m \leftarrow$ initialize weights
$U \leftarrow True$
repeat
  $\hat{G}^j_{\mathcal{X}} \leftarrow f^i(G^j_{\mathcal{X}})$, $\hat{G}^j_{x^i_p} \leftarrow s^i_p(\hat{G}^j_{\mathcal{X}})$, $\hat{G}^j_{x^i_u} \leftarrow s^i_u(\hat{G}^j_{\mathcal{X}})$
  $C \leftarrow \text{MSE}(\hat{G}^j_{\mathcal{X}}, G^j_{\mathcal{X}})$
  $l^i_p \leftarrow \text{cross\_entropy}(\hat{G}^j_{x^i_p}, G^j_{x^i_p})$
  $l^i_u \leftarrow \text{cross\_entropy}(\hat{G}^j_{x^i_u}, G^j_{x^i_u})$
  $L \leftarrow \alpha C - \lambda_p l^i_p + \lambda_u l^i_u$
  if ALFR **then**
    if U **then**
      Update $(\gamma^i_p)_m$ to *maximize* the loss L
    else
      Update $(\gamma^i)_m, (\gamma^i_u)_m$ to *minimize* the loss L
    end if
  else if UAE-PUPET **then**
    if U **then**
      Update $(\gamma^i)_m$ to *minimize* the loss L
    else
      Update $(\gamma^i_p)_m$ to *minimize* the loss l^i_p
      Update $(\gamma^i_u)_m$ to *minimize* the loss l^i_u
    end if
  end if
  $U \leftarrow$ not U
until Deadline

We provide the algorithm for the data sharing mechanism in Algorithm 3. This algorithm includes optimizations for a single iteration of a single group but can be iteratively applied to both groups using the mentioned sequence above or as visually demonstrated in Figure 3.3. In this algorithm, the variable i represents PM^i being trained by Group i using data from Group j. In the context of our scenario involving groups G1 and G2, if G1 is training $(PM^1)_m$, then $j = 2$ represents the data from G2 used for training. Similarly, if G2 is training $(PM^2)_m$, then $j = 1$. Furthermore, the subscript m denotes the iteration number and primarily applies to the function weights. While we could have included the m subscripts in the functions and losses as well, they are omitted here for readability purposes. For a detailed explanation of the notation used for various functions in the data sharing

mechanism, please refer to Table 3.1.

Notation	Description
G^i_j Example: $G^1_{\mathcal{X}}$	Raw data of Group i with feature(s) j. Raw data of G1 i.e. rows of data that belong to G1 and have $\mathcal{X}$ features.
$\hat{G}^i_j$ Example: $\hat{G}^2_{\mathcal{X}}$	Sanitized data of Group i with feature(s) j. Sanitized data of G2 i.e. rows of data that belong to G2 and have $\mathcal{X}$ features. *Exception: If $\hat{G}^i_{x^i_p}$ / $\hat{G}^i_{x^i_u}$, it represents the private/utility feature predictions from s^i_p/s^i_u.*
f^i	Function representing generator part of the privacy mechanism for Group i, (PM^i).
s^i_p, s^i_u	Function representing simulated adversary and utility provider to predict the private and utility feature of Group i, respectively. *Simulated* indicates adversaries and utilities network at the time of training.
$\gamma^i, \gamma^i_p, \gamma^i_u$	Weights of f^i, s^i_p, s^i_u, respectively.
l^i_p, l^i_u	Loss of a classifier attempting to infer the private and utility feature of Group i, respectively.

Table 3.1: Essential notations and their descriptions.

3.5 Experimental Setup

3.5.1 Dataset

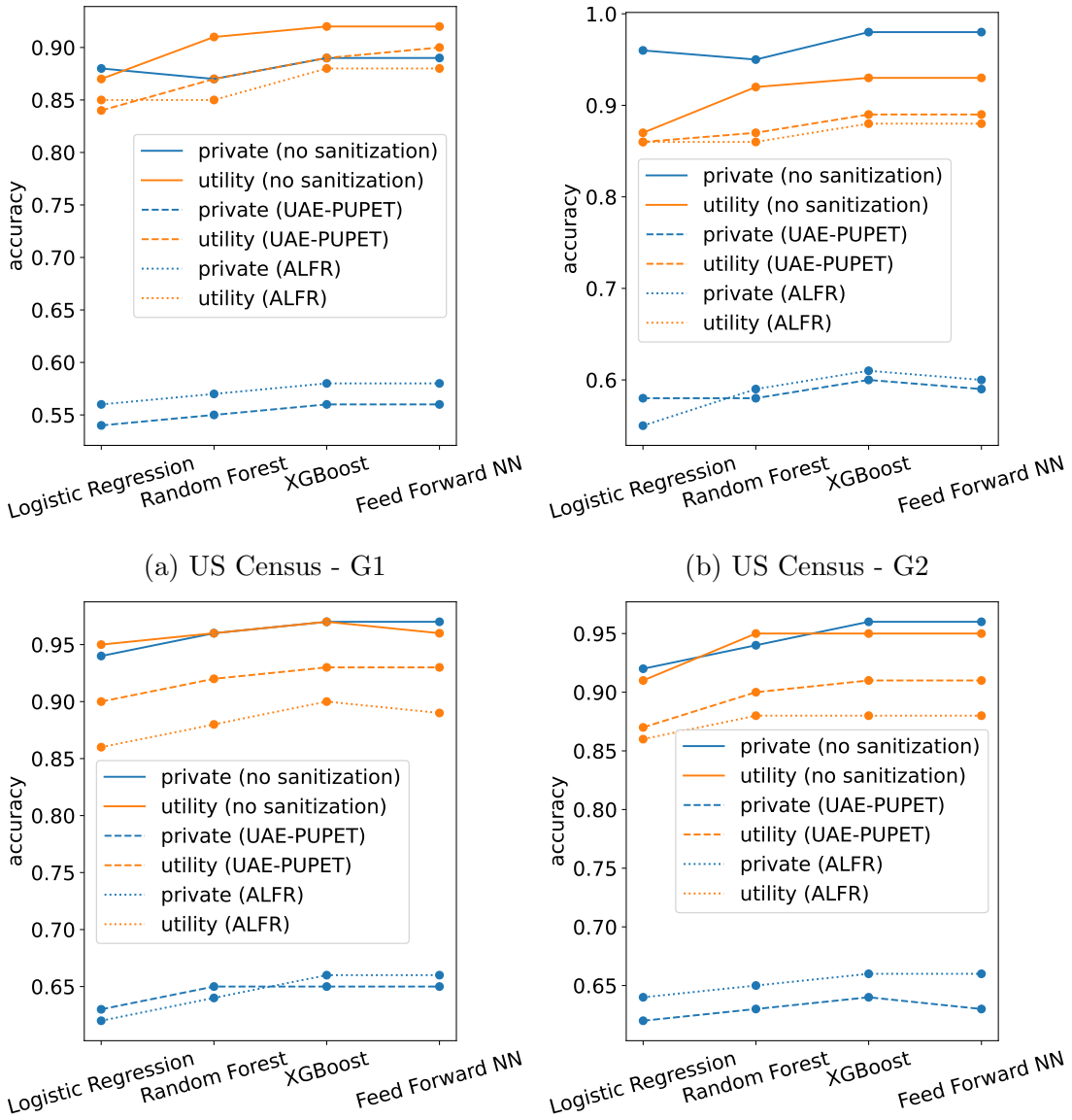
We conducted experiments using two distinct datasets: the US Census Demographic Data (referred to as *US Census*) (US Census Bureau, 2018) and a synthetic dataset generated using the scikit-learn library (Pedregosa et al., 2011). The US Census data is a 2017 American Community Survey dataset comprising 74,001 records, featuring 37 continuous variables. From the original 37 continuous variables, we selected 16 variables (e.g., male and female population, income, racial composition percentages, etc.), mirroring previous work (Sharma

et al., 2021; Mandal et al., 2022). Following preprocessing, the resulting dataset comprises 72,000 data points. The synthetic data used in our experiments also encompasses 16 continuous features and 72,000 data points. For both datasets, Groups G1 and G2 each consist of 31,000 distinct data points, while the remaining 10,000 data points are reserved as auxiliary datasets. Further details on the experiments involving auxiliary datasets are elucidated in Section 3.6.5.

Our choice to utilitze the US Census data was guided by several factors. Firstly, our requirement was for a dataset comprising four categorical features, each appropriate for addressing balanced classification task, a type of dataset that is often rare and challenging to obtain. As a result, we selected this specific dataset, which initially comprises continuous features but can be transformed into balanced classification problems for all four attributes through discretization. These attributes represent the private and utility aspects of two distinct groups. Additionally, it is worth highlighting that this dataset has previously found application in research endeavors centered on exploring the balance between privacy and utility. Furthermore, our research maintains a specific focus on tabular datasets. However, the significant emphasis on continuous input features provides us with the potential to broaden our research horizons in future work, encompassing the incorporation of image datasets or datasets containing image embeddings. Importantly, this expansion can be achieved without the necessity for supplementary post-processing after data sanitizaion, a requirement typically associated when dealing with categorical input features as mentioned in the previous chapter.

3.5.2 Private and Utility features

Regarding the US Census dataset, the private feature for Group 1 (G1) refers to the variable *Income*, whereas the of utility feature refers to the variable *Employed* which refers to the number of employed population. This choice of private and utility feature is similar to the single-group setting in (Sharma et al., 2021; Mandal et al., 2022). In our experimentation,



(a) US Census - G1

(b) US Census - G2

(c) Synthetic - G1

(d) Synthetic - G2

Figure 3.4: Performance in predicting private and utility features, for both datasets and groups, before and after data sanitization. X-axis shows different ML models tested in this study, and Y-axis represents the accuracy of prediction.

we have categorized both private and utility attributes into two balanced groups each. For the Income variable, we have established categories based on whether the income exceeds 55000 or not. Correspondingly, for the Employed variable, values equal to or below 2000 are placed into one category, and values above this threshold constitute another category.

Analogously, within Group 2 (G2), the private attribute is represented by the percentage of the white population, while the utility attribute is represented by the count of women in the population. For the white population percentage, we have delineated a category based on values surpassing 70, placing them into one class, and the rest into another class. Similarly, a class division has been established for the count of women, with values exceeding 2000 forming one category and those below constituting the other. These values have been thoughtfully selected after a thorough evaluation of the data distribution, ensuring a balanced classification scenario. The same methodology has been applied to synthetic data to attain a balanced classification setup.

3.5.3 Performance Evaluation Metrics

To assess the performance of our privacy mechanism, we utilize standard accuracy and AUROC metrics. These metrics are expected to exhibit lower values when analysts attempt to predict private features, indicating stronger privacy protection, while for utility features, higher accuracy and AUROC values are desirable.

Additionally, we leverage empirical mutual information (MI) (Gao et al., 2015) to evaluate the information-theoretic correlation between our sanitized data and the private/utility features.

Furthermore, we utilize the privacy-utility tradeoff metrics introduced in Zhang et al. (2023) where privacy leakage for attribute p is defined as follows:

$$M_p = \frac{c_a(p) - c_r(p)}{c_n(p) - c_r(p)} \tag{3.1}$$

Here, $c_n(.)$ denotes the accuracy of a classifier with raw, unsanitized data – data without undergoing any privacy mechanism. Similarly, $c_a(.)$ represents the accuracy of a classifier after data sanitization, and $c_r(.)$ signifies the accuracy of the classifier when making random guesses (in our case, 0.5, as all are balanced binary classification problems). A lower M_p value

indicates closer proximity to a random guess and, consequently, stronger privacy guarantees.

Similarly, utility performance for attribute u is quantified as follows:

$$M_u = \frac{c_a(u) - c_r(u)}{c_n(u) - c_r(u)} \tag{3.2}$$

In this case, a higher M_u value signifies less utility drop. Finally, the privacy-utility tradeoff is calculated as:

$$T = \frac{M_p}{\delta + M_u} \tag{3.3}$$

Here, $\delta = 0.0001$ is introduced to handle potential zero-division errors. A lower T value indicates a more favorable tradeoff, implying that removing the privacy feature has a less negative impact on the utility feature.

3.6 Result and Discussions

In our study, we conduct 25 independent experiments to thoroughly evaluate the effectiveness of our data sharing mechanism. We report the mean and standard deviations of the results to provide a comprehensive view of our approach's efficiency. However, in some instances, we omit reporting standard deviations for the sake of brevity, readability, or to enhance the clarity of figures. The aforementioned results correspond to the specific hyperparameter settings of $\alpha = 1$, $\lambda_u = 1$, and $\lambda_p = 0.2$, which govern the relative weighting of the competing objectives related to privacy and utility attributes.

3.6.1 Accuracy and AUROC

We utilize various ML models specifically Logistic Regression, Random Forest, XGBoost, and Feed Forward Neural Network models to first train and predict private and utility features from unsanitized data. Subsequently, we employ proposed data-sharing mechanism utilizing both ALFR and UAE-PUPET techniques to sanitize the dataset. We then utilize

the sanitized data to retrain the machine learning models from scratch and predict private and utility features. The outcomes of these diverse machine learning models on unsanitized and sanitized data are summarized in Figure 3.4, where each plot represents a distinct dataset and group. For example, Figure 3.4a presents results for Group 1 (G1) of the US Census dataset. The X-axis in these plots represents the various models used, while the Y-axis signifies the corresponding accuracy achieved by each machine learning model. Note that, each point on the plots represent mean accuracy scores of 25 experiments. Overall, for our dataset, these plots indicate that XGBoost and Feed Forward Neural Networks consistently attain the highest accuracy among the models considered for classification.

Privacy Mechanism	**Group 1 (G1)**		**Group 2 (G2)**	
	Private	**Utility**	**Private**	**Utility**
US Census Dataset				
No Privacy	0.88 ± 0.002	0.92 ± 0.003	0.98 ± 0.004	0.93 ± 0.006
ALFR	0.58 ± 0.005	0.87 ± 0.014	0.60 ± 0.011	0.88 ± 0.011
UAE-PUPET	0.55 ± 0.003	0.90 ± 0.004	0.59 ± 0.012	0.89 ± 0.009
Synthetic Dataset				
No Privacy	0.97 ± 0.003	0.96 ± 0.008	0.97 ± 0.003	0.95 ± 0.004
ALFR	0.66 ± 0.007	0.90 ± 0.009	0.66 ± 0.004	0.89 ± 0.012
UAE-PUPET	0.65 ± 0.005	0.92 ± 0.008	0.63 ± 0.006	0.92 ± 0.008

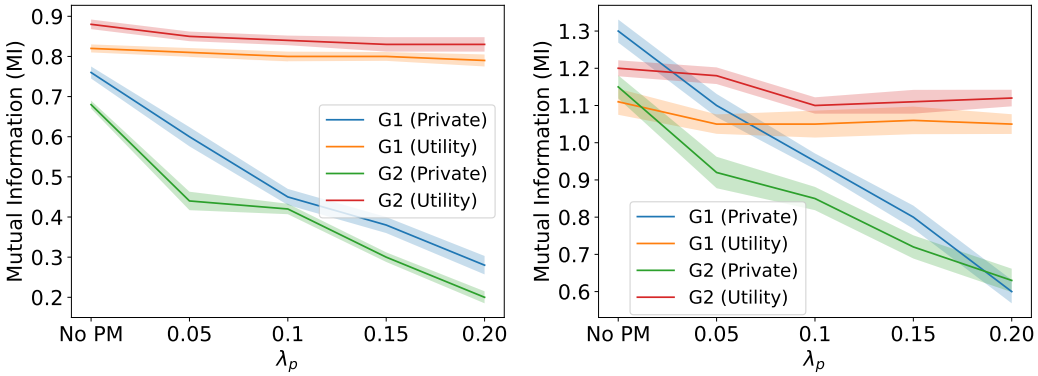
Table 3.2: Accuracy scores of predicting private and utility features before and after employing the proposed data-sharing mechanism using existing ALFR and UAE-PUPET techniques.

Among the four models employed, we identify the highest accuracy achieved as the benchmark accuarcy for predicting private or utility attributes within a particular group. We summarize these accuracy results in Table 3.2. Notably, the accuracy and AUROC scores are nearly identical, so we omit the AUROC scores for brevity. From the table, we can see that for G1 in the US Census dataset without any data sanitization (no privacy), the accuracy is 0.88 for private features and 0.92 for utility features. However, after applying the data sharing mechanism using ALFR, the accuracy for private features drops to 0.58,

and it drops to 0.55 when using the UAE-PUPET technique. In contrast, the accuracy for utility features remains high at 0.87 with ALFR and 0.90 with UAE-PUPET.

Similarly, for G2, the accuracy for private features decreases from 0.98 to 0.60 with ALFR and to 0.59 with UAE-PUPET, while the accuracy for utility features experiences a slight decrease from 0.93 to 0.88 with ALFR and 0.89 with UAE-PUPET. For results related to the synthetic dataset, please refer to Table 3.2. Overall, the data sharing mechanism proposed in this study appears to effectively safeguard both privacy and utility. Specifically, the use of UAE-PUPET in the data-sharing mechanism consistently yields lower accuracy for private features and better accuracy for utility features compared to ALFR.

3.6.2 Mutual Information



(a) US Census Dataset (b) Synthetic Dataset

Figure 3.5: Mutual Information (MI) between sanitized data derived with various λ_p values and private/utility labels. *No PM* on the X-axis indicates MI between the original(unsanitized) data and private/utility labels.

We employ mutual information as an information-theoretic measure to evaluate whether our sanitized data exhibits reduced correlation with private labels while preserving similar correlation with utility features. Figure 3.5 visually demonstrates the efficacy of our data-sharing mechanism in diminishing the mutual information between the sanitized data and private features while simultaneously maintaining the mutual information between the

sanitized data and utility features. The parameter λ_p signifies the weighted objective assigned to the privacy mechanism's loss function. A higher λ_p value, while keeping α and λ_u constant, indicates a heightened level of privacy protection. The mutual information (MI) values depicted in the figure reflect the outcomes achieved after the incorporation of UAE-PUPET's technique within our data-sharing mechanism.

3.6.3 Privacy, Utility and Tradeoffs

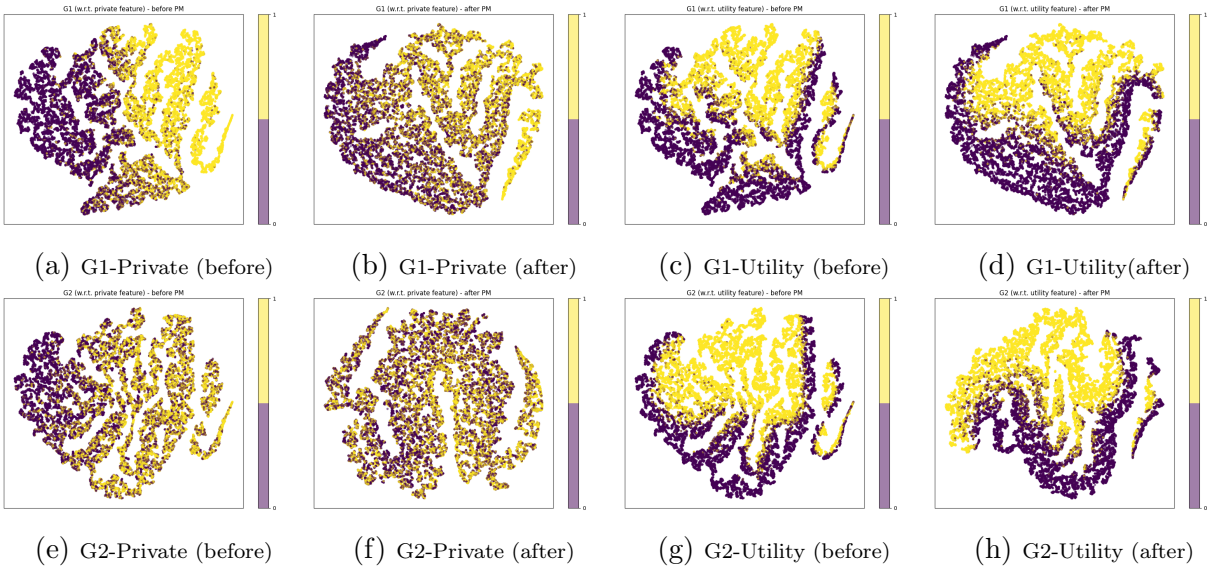
Performance Metrics	US Census Dataset		Synthetic Dataset	
	Group 1	Group 2	Group 1	Group 2
M_p	0.20, **0.15**	0.22, **0.20**	0.34, **0.31**	0.34, **0.30**
M_u	0.90, **0.95**	0.88, **0.90**	0.85, **0.91**	0.84, **0.91**
T	0.22, **0.15**	0.25, **0.22**	0.40, **0.34**	0.40, **0.32**

Table 3.3: Privacy Leakage, Utility Performance, and Privacy Utility Tradeoffs. Bold highlights the superior method between ALFR and UAE-PUPET.

Table 3.3 illustrates the result of the following metrics: Privacy Leakage (M_p), Utility Performance (M_u), and Privacy-Utility Tradeoff (T). The values in the form of a, b correspond to results obtained through the ALFR and UAE-PUPET techniques, respectively. For instance, in the US Census Dataset for Group 1, T is 0.22 (ALFR) and 0.15 (UAE-PUPET). Lower T value indicates a more favorable tradeoff implying that removing the privacy feature has a less negative impact on the utility feature. Overall, the table signifies the effectiveness of the data-sharing mechanism with both ALFR and UAE-PUPET techniques. In addition, the table highlights UAE-PUPET consistently outperforms ALFR within the data-sharing mechanism. However, the primary aim of this work is not to determine the ultimate optimal approach. Rather, it serves to introduce an innovative and real world problem formulation. Moreover, this work seeks to foster increased involvement from researchers within this field, with the aspiration that our findings can establish a standardized benchmark for forthcoming investigations.

3.6.4 Data Visualization in two-dimensions

We employ t-SNE (t-distributed Stochastic Neighbor Embedding) to create two-dimensional visualizations of both the unsanitized and sanitized US Census data concerning private and utility labels, as shown in Figure 3.6. The visualizations reveal noteworthy patterns. Before and after sanitization, the utility feature exhibits a clear separation into distinct clusters for the positive and negative classes. However, after sanitization, the private feature lacks such distinct clusters for the positive and negative categories.



(a) G1-Private (before) (b) G1-Private (after) (c) G1-Utility (before) (d) G1-Utility(after)

(e) G2-Private (before) (f) G2-Private (after) (g) G2-Utility (before) (h) G2-Utility (after)

Figure 3.6: Visualization of unsanitized and sanitized data w.r.t private and utility labels. *before* represents unsanitized data whereas, *after* represents sanitized data using UAE-PUPET technique.

3.6.5 Is privacy affected if analyst collects an auxiliary dataset?

We further explore scenarios in which an analyst possesses the capacity to acquire an additional dataset for the purpose of training various machine learning models to predict private features. Our aim is to evaluate whether the incorporation of this auxiliary dataset (G^3 data) can enhance the analyst's predictive performance. The analyst, who can also be considered an adversary, employs two distinct strategies. The first strategy involves training

exclusively on the auxiliary dataset, comprising 10,000 data points exclusive to those in Group 1 and Group 2. The second strategy entails training using a combination of both the auxiliary dataset and a sanitized dataset. We observed that both strategies yield similar predictive accuracies for the private feature, as depicted in Table 3.2, and none of the strategies demonstrate higher accuracy than what is presented in Table 3.2. This reinforces our assertion that our data sharing mechanism effectively safeguards the privacy of private features for both user groups, regardless of whether the analyst only has access to the sanitized data or can acquire an auxiliary dataset for training various predictive models.

3.7 Conclusion

In this work, we introduce a novel problem formulation that focuses on balancing the tradeoff between privacy and utility within two distinct user groups. Notably, the third-party responsible for sanitizing the dataset has no access to any auxiliary dataset and is not required to annotate data separately for different user groups. Instead, the proposed data sharing mechanism leverages alternate group data. Additionally, analysts are restricted to utilizing only the sanitized data for training their predictive models. Our proposed data sharing mechanism can be seamlessly integrated with existing privacy techniques. In our work, we demonstrate the applicability of the data sharing mechanism in conjunction with ALFR and UAE-PUPET. Both of these techniques, when incorporated into the data sharing mechanism, have proven effective in reducing the analyst's accuracy in predicting private features while maintaining a high level of accuracy in predicting utility features. Furthermore, our experiments indicate that even when analysts gather auxiliary datasets, the privacy of both user groups remains intact.

The primary objective of this work is to introduce a novel real-world problem formulation, addressing a gap that exists despite the substantial research efforts dedicated to the intricate landscape of the privacy-utility tradeoff. Moreover, the problem formulation

and the proposed solution have the potential for broader applications in mitigating biases, enhancing fairness, and rectifying discrimination in machine learning models, especially in scenarios involving multiple groups.

Chapter 4

Privatization Mechanisms to Attain Privacy and Utility Tradeoffs in Multi-Party Environments

4.1 Introduction

In the preceding chapter, Chapter 3, we delved into a scenario featuring two distinct user groups, facilitated by a trusted third-party service provider. This intermediary enjoyed the confidence of both groups, thereby receiving the original data from each. Within this framework, the task of training privacy mechanisms was notably simplified, given the third party's possession of clean data and labels.

Building upon this groundwork, in this chapter, we embark on an extension of the aforementioned work. Here, we remove the assumption of a trusted third-party service provider who possess clean data and labels. Additionally, we confront a scenario where the data housed in the final database comprises not only perturbed data but also perturbed labels. In response to this paradigm shift, we introduce three distinct training mechanisms aimed at extending the existing privacy mechanisms to address this new challenge. For

simplicity, we again consider a scenario with two groups of users who have different private and utility features. Additionally, we explore different methods of training our classifiers, which allows for the inference of utility features with high accuracy while effectively hiding the private features.

The forthcoming section elucidates a detailed problem formulation, laying the foundation for our exploration into innovative solutions for privacy-preserving data processing amidst a decentralized and less-trusting data environment.

4.2 Problem Formulation

Consider a setting where we have k individuals who are seeking to share their data in a database in order to gain a certain level of utility while also maintaining privacy regarding certain feature or group of features. The requirements for utility and privacy may vary among these users, and what one user considers private information may not be seen as such by another. Additionally, the desired utility features may also differ among the users. In order to model the aforementioned scenario, we can consider a dataset G with k rows and n features, where the n features are represented by $\mathcal{X} = \{x_1, x_2, x_3, \cdots x_n\}$. Each row represents the data of a single user. Assuming that each user has one private feature and one utility feature from a set of n features, there are a total of $\binom{n}{2}$ private and utility feature choices. The purpose of this paper is not to model such large number of choices. Instead, for simplicity, we divide the k users into two equal groups – *Group 1 (G1)* and *Group 2 (G2)*. G1 consists of users with one set of private and utility feature represented by x_p^1 and x_u^1, respectively. Similarly, G2 consists of users with another set of private and utility feature represented by x_p^2 and x_u^2, respectively. Note that, $\{x_p^1, x_u^1, x_p^2, x_u^2\} \subset \mathcal{X}$. We note here that G1 does not disclose x_p^1 and x_u^1, and G2 does not disclose x_p^2 and x_u^2. Nevertheless, these features can be inferred from the other disclosed features, by virtue of correlation. Such correlation can, in principle, be learned from the data disclosed by other groups –

for instance, we expect that x_p^1 and x_u^1 are usually disclosed by G2 – albeit through some privacy mechanism – and x_p^2 and x_u^2 are disclosed by G1, allowing an adversary to learn how to infer x_p^i and x_u^i from $\mathcal{X} \setminus \{x_p^i,\ x_u^i\}$.

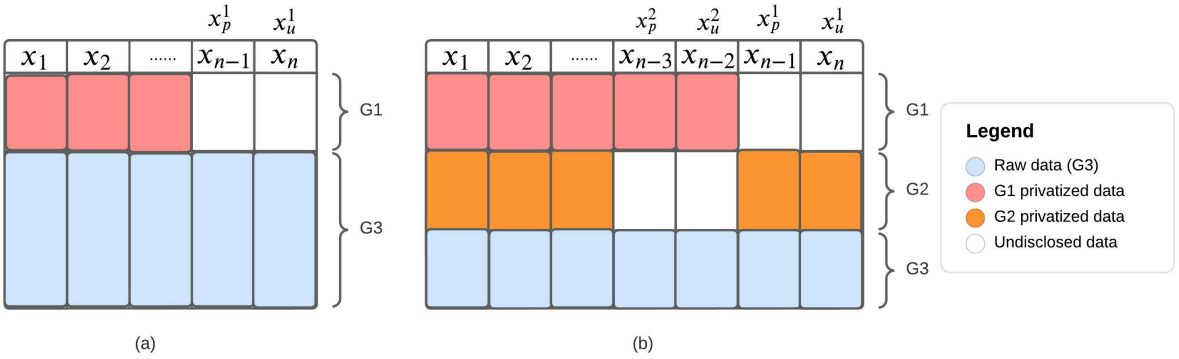


Figure 4.1: (a) Single Party Privacy Utility Tradeoff setting, (b) Multi-Party Privacy Utility Tradeoff setting.

4.2.1 Single Party privacy-utility tradeoff

In the context of single party privacy-utility tradeoff problem, there is a single group (G1) with private and utility features, x_p^1 and x_u^1 respectively. The users in this group disclose the set $\mathcal{A} = \mathcal{X} \setminus \{x_p^1, x_u^1\}$ of features, where they wish the utility provider to be able to infer x_u^1 feature from the disclosed $\mathcal{A}$ features, whereas the adversaries to not be able to infer x_p^1 feature from the disclosed $\mathcal{A}$ features. To preserve some generality, we assume the existence of a group G3 representing users with no privacy requirements, who hence disclose their entire raw data $\mathcal{X}$ (refer to Fig. 4.1 (a)). The *threat model* refers to adversaries/attackers having access to a large amount of G3 data, and therefore capable of training classifiers using this raw data from G3 to infer the private features of G1. In order to decrease the capability of the adversary's model to correctly estimate the private feature, and to maintain the capability of the utility provider's model to correctly estimate the utility feature, G1 users firstly employ a privacy mechanism (PM) and generate privatized data which is then disclosed publicly. The objective of the PM, for example UAE-PUPET, is to implicitly

minimize $I(\hat{G}^1_{\mathcal{A}}\ ;\ G^1_{x^1_p})$ and *maximize* $I(\hat{G}^1_{\mathcal{A}}\ ;\ G^1_{x^1_u})$, where $I(X;Y)$ represents the mutual information between two vectors X and Y, $\hat{G}^i_J$ represents privatized data of Group i with feature set J, and G^i_J represents the raw data of Group i with feature set J. To fulfill this objective, the PM is trained using only the data from G3 such that the mutual information (I) between $G^3_{\mathcal{A}}$ and $G^3_{x^1_p}$ is minimized and the mutual information between $G^3_{\mathcal{A}}$ and $G^3_{x^1_u}$ is maximized. Once this PM has been trained using the G3 data, it takes the data that needs to be privatized i.e. G1 data, excluding x^1_p and x^1_u features as they are not disclosed publicly, and generates privatized data $\hat{G}^1_{\mathcal{A}}$. Ideally, the adversary should be unable to accurately infer $G^1_{x^1_p}$ with high accuracy, while the utility provider should be able to infer $G^1_{x^1_u}$ with high accuracy. The single-party case, if extended to two different user groups with different private and utility features, still assumes adversary and utility provider to have access to large amount of G3 data.

The single-party scenario assumes the presence of a large amount of G3 data. It assumes that all users training their privacy mechanisms, as well as adversaries and utility providers training their classifiers, use only the G3 data, even when there are multiple user groups with different private and utility features. However, as more users adopt privacy mechanisms, it may become challenging to have a sufficient amount of G3 data available. In such cases, it becomes necessary to utilize privatized data from other groups to train the privacy mechanism. Likewise, the adversaries and utility providers may also require privatized data from other groups to train their classifiers. This is referred to as the multi-party privacy utility tradeoff.

4.2.2 Multi-Party privacy-utility tradeoff

The multi-party privacy-utility tradeoff scenario involves multiple groups of users, each with their own private and utility features. *In this work, we consider the simplest such scenario, consisting of only two groups G1 and G2, along with the privacy-indifferent group G3. Nevertheless, our methodology and results are easily generalizable to more groups.*

G1 and G2 train their own privacy mechanisms (PM^1 and PM^2) internally using simulated adversaries (s_p^1, s_p^2) and utility providers (s_u^1, s_u^2) respectively. These groups then share only the privatized data, represented by $\hat{G}_{\mathcal{A}}^1$ and $\hat{G}_{\mathcal{B}}^2$, respectively, with the database where $\mathcal{B} = \mathcal{X} \setminus \{x_p^2, x_u^2\}$ set of features. The real adversary (r_p^1, r_p^2) and utility providers (r_u^1, r_u^2) therefore have access to the privatized data of G1 and G2, and raw data of G3 to infer private and utility features (refer to Fig. 4.1 (b)), unlike the single-party privacy setting where the only data used by adversaries and utility providers to train their classifiers was the raw data G3. Note that r_p^i, r_u^i refer to the functions representing classifiers to predict the private and utility feature of Group i, respectively, after privatized data has been disclosed. That is, to infer the private and utility feature of G1 as a supervised learning problem, adversaries and utility provider should be trained using the data that G2 shares i.e. $\hat{G}_{\mathcal{B}}^2$ and G3 data. Similarly, to infer private and utility feature of G2 as a supervised learning problem, classifiers must be trained from the privatized data of G1 i.e. $\hat{G}_{\mathcal{A}}^1$ and G3 data. It should be noted that adversaries and utility providers have the flexibility to employ different tactics for training their classifiers, which will be elaborated in Section 4.3.4.

Formally, the multi-party privacy-utility tradeoff refers to maintaining the privacy of both user groups while still allowing for the useful analysis of the data to fulfill the following objectives:

- *minimize* $I(\hat{G}_{\mathcal{A}}^1 \; ; \; G_{x_p^1}^1)$ and $I(\hat{G}_{\mathcal{B}}^2 \; ; \; G_{x_p^2}^2)$
- *maximize* $I(\hat{G}_{\mathcal{A}}^1 \; ; \; G_{x_u^1}^1)$ and $I(\hat{G}_{\mathcal{B}}^2 \; ; \; G_{x_u^2}^2)$.

4.2.3 A trivial solution to a pathological case

Minimizing mutual information (I) between disclosed features and private features and maximizing I between the disclosed features and utility features can be trivial if there is no G3 present i.e. G3 contains zero users (rows). Specifically, our trivial solution can lead to no loss of information in utility and perfect protection i.e. zero mutual information, about

private features.

Theorem 1. *The two-group scenario supports a solution that achieves perfect privacy and zero loss of utility.*

Proof. To achieve such results, both groups, G1 and G2 agree on a setting where they do not disclose their own private features or those of the other group in a shared database, G. Instead, G1 can share the features $\mathcal{X} \setminus \{x_p^1, x_u^1, x_p^2\}$ and G2 can share the features $\mathcal{X} \setminus \{x_p^2, x_u^2, x_p^1\}$. This arrangement allows for no loss in mutual information about the utility features because no distortion is applied to the available data. Note here that the fact that G1 does not share x_p^2 incurs no loss for G1, as no training data is available to a utility provider that contains x_p^2, so this particular feature cannot be used to increase G1's utility. This is not so when a group G3 is available. The solution offers perfect privacy of the private features because no adversary can train a classifier to infer the private features of either group, as no information about these features is disclosed, and hence no training labels are available. □

The trivial solution for addressing the above pathological multi-party privacy-utility tradeoff problem has little applicability in practice. This is because such a pathological multi-party privacy-utility tradeoff problem is not very realistic, as it assumes that the private labels can only be obtained from one of the two colluding, privacy-concerned user groups. The pathological scenario also emphasizes the significance of the available size of G3 data. Therefore, our problem becomes intriguing when examining the progression of the multi-party privacy utility tradeoff for varying sizes of G3 data.

4.3 Methodology

Since there is no mutually trusted third-party service provider acknowledged by both groups, employing a singular adversarial mechanism to sanitize the datasets for both parties is

unfeasible. Therefore, we utilize distinct privacy mechanisms for the two different groups, i.e. two adversarial training mechanisms referred to as PM^1 and PM^2. PM^1 is a privacy mechanism that is used to privatize G1 data and PM^2 is a privacy mechanism that is used to privatize G2 data. To demonstrate the compatibility of our suggested training approaches with existing privacy mechanisms, we again employ two distinct privacy mechanisms: ALFR (Edwards and Storkey, 2016) and UAE-PUPET (Mandal et al., 2022). These mechanisms were initially designed to tackle the balance between privacy and utility in single-party scenarios.

ALFR and UAE-PUPET

ALFR (Adversarial Learned Fair Representations) (Edwards and Storkey, 2016) and UAE-PUPET (Uncertainty Autoencoder based Privacy and Utility Preserving end-to-end Transformation) (Mandal et al., 2022) consist of a neural network architecture that includes a generator part and a discriminator part. The generator part consists of a generative model composed of an encoder-decoder pair, while the discriminator part consists of classifiers that predict the private and utility features. ALFR and UAE-PUPET both were proposed to solve single-party privacy-utility tradeoffs. Recall that in the single-party setting there is large group, G3, which has no privacy requirements and thus shares raw data. G3 data is used for training these privacy mechanisms and data from G1 needs to be privatized after these PMs are trained. The generator in its most general form is represented as a function f^1 that takes training input as $G^3_{\mathcal{A}}$ and generates $\hat{G}^3_{\mathcal{A}}$ i.e. $\hat{G}^3_{\mathcal{A}} = f^1(G^3_{\mathcal{A}})$. ALFR uses a standard autoencoder (AE) whereas UAE-PUPET uses an Uncertainty Autoencoder (UAE) (Grover and Ermon, 2018) as a generator. UAE incorporates randomness and does not impose any assumptions regarding the Gaussian or any prior distribution on the latent variable or bottleneck, unlike the Variational Autoencoder (Kingma and Welling, 2022). The degree of randomness can be adjusted using a hyperparameter specified by the user.

The discriminators used in both ALFR and UAE-PUPET employ neural networks r^1_p

and r_u^1, with three hidden layers each, which are responsible for predicting the private and utility features, respectively.

Although ALFR and UAE-PUPET were initially introduced to address privacy-utility tradeoffs in single-party scenarios, we present three training approaches: *Individual* approach, *Data Sharing* approach, and *Model Sharing* approach to handle multi-party privacy-utility tradeoffs. It is crucial to clarify that in this context, the term *privacy mechanism* encompasses the training approach, while PM^1 and PM^2 denote the specific adversarial machine learning models responsible for privatizing the data of their respective groups, G1 and G2.

Note: In this chapter, we revisit the discussion on ALFR and UAE-PUPET. This revisit is motivated by potential disparities in notations between this chapter and Chapter 3. Moreover, we address a slightly modified data sharing approach in the context of this new problem formulation, distinct from that in previous chapter, Chapter 3.

4.3.1 Individual Approach

In this approach, two trusted parties are responsible for collecting their respective group's data, G1 and G2. These parties collect the data for the full set of features, $\mathcal{X}$. PM^1 and PM^2 are trained using a portion β of G1 and G2 data, respectively. Note that in this approach PM^1 and PM^2 do not make use of G3 data. The generator (f^1, f^2) takes $G^1_{\mathcal{A}}$ and $G^2_{\mathcal{B}}$ as input for G1 and G2, while the discriminator receives the $G^1_{x_p^1}, G^1_{x_u^1}$ and $G^2_{x_p^2}, G^2_{x_u^2}$ data as labels for G1 and G2, respectively. The discriminator part consists of simulated adversaries and utility providers, whose functions are represented by s_p^i, s_u^i where s_p^i learns to predict the private feature x_p^i, and s_u^i learns to predict the utility feature x_u^i, of group G^i. The term *simulated* adversaries and utility providers is utilized because, during evaluation or real-life usage, there may exist other adversaries and utility providers who can adopt distinct tactics for inferring private and utility features, respectively. Detailed explanations of these alternative approaches are provided in Section 4.3.4. Once PM^1 and PM^2 are trained, they

use the entire G1, G2 data to produce privatized versions of G1 and G2, respectively, i.e. $\hat{G}^1_{\mathcal{A}} = f^1(G^1_{\mathcal{A}})$ and $\hat{G}^2_{\mathcal{B}} = f^2(G^2_{\mathcal{B}})$ and disclose them in the database G publicly. Note that the training of each of PM^1 and PM^2 occurs in an *enclave*.

The individual approach is presented in Algorithm 4. In this algorithm, γ^i represents the neural network weights for the generator function f^i, whereas γ_{P^i} and γ_{U^i} correspond to the neural network weights for the adversary and utility provider functions s^i_p and s^i_u, respectively. The variable i indicates Group i in the algorithm, and can be generalized to not only two groups G1 and G2 but can be used by arbitrary number of groups. Furthermore, the inclusion of ALFR or UAE-PUPET in the algorithm is intended to demonstrate that our training approach is compatible with both of these techniques. It should be noted that alternative techniques, apart from ALFR and UAE-PUPET, can also be employed. The same applies to the other two approaches mentioned in the subsequent sections.

4.3.2 Data Sharing Approach

The data sharing approach is an extension of the individual approach where G1 and G2 internally share their privatized data, $\hat{G}^1_{\mathcal{A}}$ and $\hat{G}^2_{\mathcal{B}}$, with each other. If G1 anticipates that G2 will publicly disclose $\hat{G}^2_{\mathcal{B}}$, G1 can use this information to retrain its privacy mechanism, PM^1_m, using $\hat{G}^2_{\mathcal{B}}$ and $G^3_{\mathcal{X}}$. G1 trains PM^1_m using simulated adversaries and utility providers in order to closely emulate the behavior of real adversaries and utility providers, who will only have access to $\hat{G}^2_{\mathcal{B}}$ and $G^3_{\mathcal{X}}$ to train their classifiers. This allows G1 to further privatize its data, generating $(\hat{G}^1_{\mathcal{A}})_m = (f^1)_m(G^1_{\mathcal{A}})$, where subscript m denotes the index of the corresponding iterative process and f^1 represents the generator part of PM^1, which it then shares with G2 ($m = 1$ represents the start of the iterative process). The weights for the generators are saved at each iteration m. G2 can then use $(\hat{G}^1_{\mathcal{A}})_m$ and $G^3_{\mathcal{X}}$ to retrain its own privacy mechanism, PM^2_m, and generate privatized data, $(\hat{G}^2_{\mathcal{B}})_m = (f^2)_m(G^2_{\mathcal{B}})$. G1 can then use this new privatized data to retrain its privacy mechanism PM^1_{m+1}, and G2 can then retrain its privacy mechanism PM^2_{m+1} using the updated privatized data from G1.

Algorithm 4 Individual Approach

$\gamma^i, \gamma_{P^i}, \gamma_{U^i} \leftarrow$ initialize weights
$U \leftarrow True$ ▷ Boolean indicating whether to update weights γ^i or $(\gamma_{P^i}, \gamma_{U^i})$
repeat
  $\hat{G}^i_{\mathcal{V}} \leftarrow f^i(G^i_{\mathcal{V}})$, $\hat{G}^i_{x^i_p} \leftarrow s^i_p(\hat{G}^i_{\mathcal{V}})$, $\hat{G}^i_{x^i_u} \leftarrow s^i_u(\hat{G}^i_{\mathcal{V}})$
  $C \leftarrow MSE(\hat{G}^i_{\mathcal{V}}, G^i_{\mathcal{V}})$ ▷ Reconstruction loss for generator
  $l^i_p \leftarrow cross\_entropy(\hat{G}^i_{x^i_p}, G^i_{x^i_p})$ ▷ Cross entropy loss for the adversary network
  $l^i_u \leftarrow cross\_entropy(\hat{G}^i_{x^i_u}, G^i_{x^i_u})$ ▷ Cross entropy loss for the utility network
  $L \leftarrow \alpha C - \lambda_p l^i_p + \lambda_u l^i_u$ ▷ Joint Loss
  if ALFR **then**
    if U **then**
      Update weights (γ_{P^i}) to *maximize* the loss L
    else
      Update weights $(\gamma^i, , \gamma_{U^i})$ to *minimize* the loss L
    end if
  else if UAE-PUPET **then**
    if U **then**
      Update weights (γ^i) to *minimize* the loss L
    else
      Update weights (γ_{P^i}) to *minimize* the loss l^i_p
      Update weights (γ_{U^i}) to *minimize* the loss l^i_u
    end if
  end if
  $U \leftarrow$ not U
until Deadline

This process can be repeated until T iterations i.e. $m \in \{1 \cdots T\}$ (refer to Fig. 4.2). The iteration with the desired results is then used to generate the final privatized data, which is publicly disclosed. The final disclosed data from G1 is represented by $\hat{G}^1_{\mathcal{A}}$, and the final disclosed data from G2 is represented by $\hat{G}^2_{\mathcal{B}}$ (iteration subscript is no longer required). It is important to note that when training PM^1_m, the generator should be able to take all the variables that G1 intends to disclose as input. However, the training data from $\hat{G}^2_{\mathcal{B}}$ does not include the features x^2_p and x^2_u. Therefore, the first step is to impute (using neural networks) these values using data from G3, and then use all the $\mathcal{A}$ features as input to the generator, and x^1_p, x^1_u features as input to the discriminator. Similarly, G2 should also perform the necessary imputation similar to G1.

Algorithm 5 Data Sharing Approach

$(\gamma^i)_m, (\gamma_{P^i})_m, (\gamma_{U^i})_m \leftarrow$ initialize weights
$U \leftarrow True$ ▷ Boolean indicating whether to update weights $(\gamma^i)_m$ or $((\gamma_{P^i})_m, (\gamma_{U^i})_m)$
repeat
  $\hat{G}^j_{\mathcal{V}} \leftarrow f^i(G^j_{\mathcal{V}})$, $\hat{G}^j_{x^i_p} \leftarrow s^i_p(\hat{G}^j_{\mathcal{V}})$, $\hat{G}^j_{x^i_u} \leftarrow s^i_u(\hat{G}^j_{\mathcal{V}})$
  $C \leftarrow MSE(\hat{G}^j_{\mathcal{V}}, G^j_{\mathcal{V}})$ ▷ Reconstruction loss for generator
  $l^i_p \leftarrow cross\_entropy(\hat{G}^j_{x^i_p}, G^j_{x^i_p})$ ▷ Cross entropy loss for the adversary network
  $l^i_u \leftarrow cross\_entropy(\hat{G}^j_{x^i_u}, G^j_{x^i_u})$ ▷ Cross entropy loss for the utility network
  $L \leftarrow \alpha C - \lambda_p l^i_p + \lambda_u l^i_u$ ▷ Joint Loss
  if ALFR **then**
    if U **then**
      Update weights $((\gamma_{P^i})_m)$ to *maximize* the loss L
    else
      Update weights $((\gamma^i)_m, , (\gamma_{U^i})_m)$ to *minimize* the loss L
    end if
  else if UAE-PUPET **then**
    if U **then**
      Update weights $((\gamma^i)_m)$ to *minimize* the loss L
    else
      Update weights $((\gamma_{P^i})_m)$ to *minimize* the loss l^i_p
      Update weights $((\gamma_{U^i})_m)$ to *minimize* the loss l^i_u
    end if
  end if
  $U \leftarrow$ not U
until Deadline

The optimizations to be carried out in the data sharing approach are outlined in Algorithm 5. In this algorithm, the variable i denotes PM^i being trained by Group i using Group j's data. This framework can be scaled to accommodate an arbitrary number of groups. In our scenario involving groups G1, G2, and G3, where G1 and G2 need to privatize their data, if G1 is training PM^1, then $j = 2, 3$ represents the data from G2 and G3 used for training. Similarly, if G2 is training PM^2, then $j = 1, 3$. Additionally, in our setting of groups G1, G2, and G3, $\mathcal{V} = \mathcal{A}$ represents the set of features when $i = 1$, and $\mathcal{V} = \mathcal{B}$ represents the set of features when $i = 2$. Furthermore, the subscript m denotes the iteration number and is mainly applied to the function weights. Although we could have included the m subscripts in the functions and losses as well, they are omitted here for the

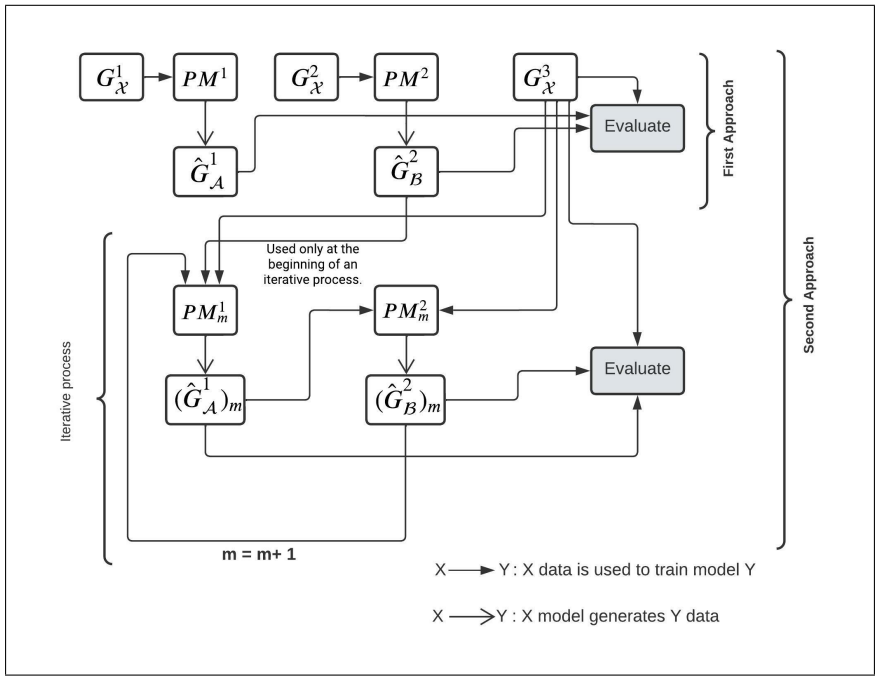


Figure 4.2: Flowchart representing the first (individual) and second (data-sharing) approach. In the individual approach, two privacy mechanisms, PM^1 and PM^2, are trained using $G^1_{\mathcal{X}}$ and $G^2_{\mathcal{X}}$, respectively. In the data-sharing approach, PM^1_m and PM^2_m are sequentially trained on each other's privatized data along with G3 data.

sake of readability.

4.3.3 Model Sharing Approach

In this approach, instead of G1 and G2 sharing their privatized data with each other, they share their privacy mechanisms, PM^1 and PM^2, respectively. The process for attaining a privacy-utility tradeoff between two groups/parties is as follows:

1. PM^1 and PM^2 are trained using data from $G^1_{\mathcal{X}}$ and $G^2_{\mathcal{X}}$, respectively. Instead of sharing privatized data from these privacy mechanisms, G1 and G2 share PM^1 and PM^2 with each other.

2. G1, who starts the iterative process, uses the generator of PM^2 shared by G2 to generate privatized data, $\hat{G}^1_{\mathcal{B}}$, based on its own data $G^1_{\mathcal{B}}$. Note that $\mathcal{B}$ features are

present in this privatized data. Using $\hat{G}^1_{\mathcal{B}}$ and data from the third group, G3, G1 can retrain its privacy mechanism, PM^1_m, which is then shared with G2 to continue the iterative process of model sharing. Here, $m = 1$ represents the start of the iterative process. Importantly, as $\hat{G}^1_{\mathcal{B}}$ is a privatized version of G1's own data, it can replace the privatized versions, $\hat{x}^1_p$ and $\hat{x}^1_u$, with the original x^1_p and x^1_u features. This is a key difference between the iterative process of the model sharing approach and the data sharing approach, as it allows for optimization using true labels (raw labels) whereas in the data sharing approach only privatized labels are present, except for G3. Additionally, any missing features can be compensated by replacing them with G1's raw data for that feature.

3. G2, who now has PM^1_m, uses its own data and PM^1_m to generate $\hat{G}^2_{\mathcal{A}}$. G2 can then retrain its privacy mechanism, PM^2_m, using $\hat{G}^2_{\mathcal{A}}$ and data from G3, and either share it with G1 to continue the iterative process or generate privatized data to disclose to the public. As with G1, from $\hat{G}^2_{\mathcal{A}}$, $\hat{x^2_p}$ and $\hat{x^2_u}$ can be replaced with the original x^2_p and x^2_u features.

4. Steps 2 and 3 can be repeated T times. The results from each iteration $m \in \{1 \cdots T\}$ is recorded, and the iteration with the most favorable results is used to generate the final privatized data. This data is then made available to the public. The final data from G1 is represented by $\hat{G}^1_{\mathcal{A}}$, and the final data from G2 is represented by $\hat{G}^2_{\mathcal{B}}$.

Remarks

In the data-sharing and the model sharing approaches, we progress through the iterative process, in the hope that a new iteration makes privacy and utility guarantees better. However, it might be the case that the result in, say, the second iteration ($m = 2$) is better than the last iteration T. In this case the data that is disclosed is the one using the generator's weights $(\gamma^i)_m$ where m = 2.

Moreover, it should be noted that both ALFR and UAE-PUPET methods do not offer formal guarantees but instead rely on heuristic solutions. Similarly, our three proposed training approaches also adopt a heuristic approach without formal guarantees, but instead provide practical solutions. These heuristics indicate that we do not provide formal guarantees of convergence, although we do acknowledge that the solution to the min-max optimization problem represents a fixed point in the process. We recognize that training the min-max optimization poses challenges, and the inclusion of additional groups further complicates the process. Therefore, future research should aim to develop theoretical frameworks or alternative strategies to enhance stability.

4.3.4 Evaluation / Threat model

In this section, we describe three methods by which the adversaries and utility providers can potentially train their classifiers to infer private and utility features, respectively. Note that these methods are used by the adversaries and utility providers only after the privacy mechanisms have completed the privatization process in enclaves by either of the three approaches and disclose their privatized data to the database G. These classifiers are referred to as the *real adversary* r_p^1, r_p^2, and *real utility provider* r_u^1, r_u^2.

G3 only (Evaluation / Threat 1)

In this method, all of the classifiers $r_p^1, r_p^2, r_u^1, r_u^2$ are trained using only the G3 data. Specifically, r_p^1 and r_u^1 are trained using $G_{\mathcal{A}}^3$ to predict $G_{x_p^1}^3$ and $G_{x_u^1}^3$, respectively. After training is complete, r_p^1 and r_u^1 are tested to predict $G_{x_p^1}^1$ and $G_{x_u^1}^1$ from $\hat{G}_{\mathcal{A}}^1$. Similarly, r_p^2 and r_u^2 are trained using $G_{\mathcal{B}}^3$ to predict $G_{x_p^2}^3$ and $G_{x_u^2}^3$, respectively. After training is complete, r_p^2 and r_u^2 are tested to predict $G_{x_p^2}^2$ and $G_{x_u^2}^2$ from $\hat{G}_{\mathcal{B}}^2$.

Using all available data without imputation (Evaluation / Threat 2)

In this method, the classifiers r_p^1 and r_u^1 are trained using the G3 data and the privatized data $\hat{G^2_{\mathcal{B}}}$. While G3 shares all $\mathcal{X}$ features, G2 only shares the $\mathcal{B}$ features. To make use of data from both groups, we concatenate the two data sets with $\mathcal{B}$ features, ignoring features that G2 does not share. Generally, $G_{\mathcal{B}}^{w,z} = \hat{G}_{\mathcal{B}}^{w} \cup G_{\mathcal{B}}^{z}$ represents the concatenation of w and z groups of data with $\mathcal{B}$ features. Then, r_p^1 and r_u^1 are trained on $G_{\mathcal{B}}^{2,3}$ to predict $G_{x_p^1}^{2,3}$ and $G_{x_u^1}^{2,3}$, respectively. After training, r_p^1 and r_u^1 are tested on $\hat{G}_{\mathcal{A}}^1$ to predict $G_{x_p^1}^1$ and $G_{x_u^1}^1$, respectively. In a similar way, classifiers r_p^2 and r_u^2 are trained using the G3 data and the privatized data $\hat{G}_{\mathcal{B}}^1$, and are tested on $\hat{G}_{\mathcal{B}}^2$ to predict $G_{x_p^2}^2$ and $G_{x_u^2}^2$, respectively.

Using all available data with imputation (Evaluation / Threat 3)

In this method, the process for training the classifiers r_p^1 and r_u^1 using the G3 data and privatized data $\hat{G}_{\mathcal{B}}^2$ is similar to the previous method. However, before concatenating the two data sets, $\hat{G}_{\mathcal{B}}^2$ and $G_{\mathcal{X}}^3$, all $\mathcal{X}$ features are included rather than just the $\mathcal{B}$ features. To accomplish this, neural network models are trained to impute the missing features using the G3 data. The classifiers r_p^2 and r_u^2 are trained in a similar manner using the G3 data and the privatized data $\hat{G}_{\mathcal{A}}^1$.

4.4 Experiment and Results

We conduct experiments using two datasets: the US Census Demographic Data (US Census) (US Census Bureau, 2018) and a synthetic dataset generated using scikit-learn (Pedregosa et al., 2011). The US Census Demographic Data is a 2017 American Community Survey dataset comprising 74,001 records, featuring 37 continuous variables. We select 16 of these variables (eg. population of men, women, income, percentage of certain races, etc), similar to (Sharma et al., 2021; Mandal et al., 2022). After preprocessing, the resulting dataset consists of over 72,000 data points, which we divide into three groups. The synthetic data used in

the experiments also includes 16 continuous features and 72,000 data points. G1 and G2 each contain 31,000 distinct data points. As previously mentioned, the pathological scenario underlines the importance of the available size of G3 data. Moreover, it is clear that if G3 is very large, there is no need for either the privacy mechanism or the adversary and utility provider to use privatized data to train their respective models; they can simply use the G3 data only. Consequently, in our experiments, we employ varying sizes $[0, 10, 100, 1000, 10000]$ of G3 data to examine the impact of different data sizes on the multi-party privacy process. This analysis also enables us to determine which of the three techniques outlined in Section 4.3.4 should be employed by real adversaries and utility providers to achieve the highest accuracies in inferring private and utility features, respectively.

We aim to simulate a situation where internet interactions generate a vast amount of data, such as engagement time, link clicks, etc. Instead of determining exact metrics, such as the duration of time spent on a post, the adversaries and utility providers usually employ machine learning models to learn categories of these variables in order to categorize users. In order to prevent correct inference of private features and enable accurate inference of utility features into relevant categories, the privacy providing parties disclose the private and utility features in categories rather than as continuous variables. Note that the privacy mechanism takes all the features as continuous variables during training, except for that particular group's private and utility labels which are provided as categories to the discriminator. However, while disclosing data, the mechanism goes through post processing to disclose the private and utility features of the other group as categorical variables.

In our experiment we divide private and utility features to two balanced categories each. For the US Demographic Census data, the private feature of G1 is the Income variable, while the utility feature is the variable Employment. Similarly, the percentage of white population, and no. of women in the population represent the private and utility features of G2, respectively. We also created similar divisions for the synthetic data to achieve a balanced classification problem. The outcomes of our experiments employing various

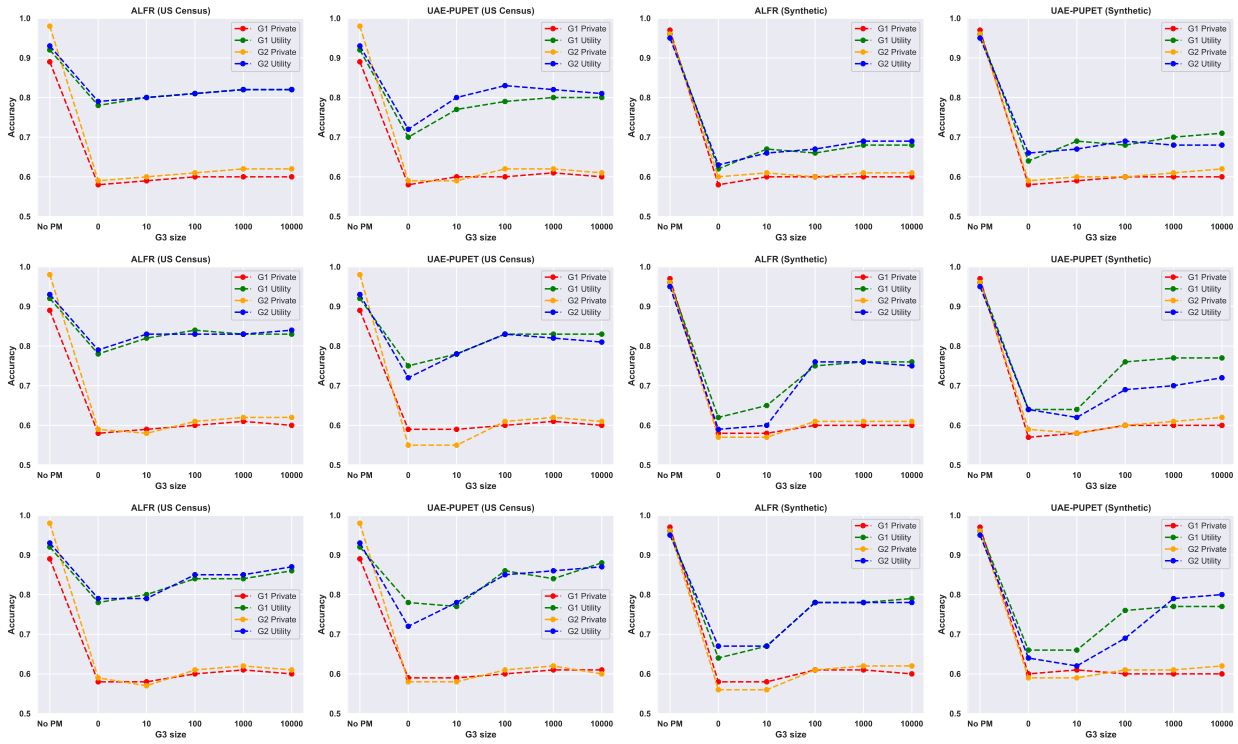
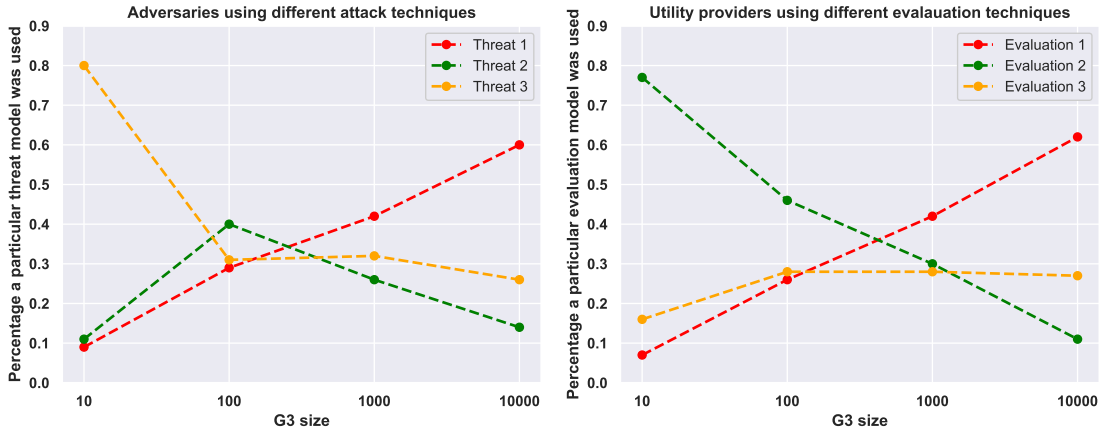


Figure 4.3: Results of all three training approaches. The first row represent results from Individual approach, second row represents results from Data Sharing approach and third row represents results from Model sharing approach. The X-axis represent size of G3 and Y-aixs represent inference accuracy. Note that No PM refers to the baselines when no privacy mechanism is used.

approaches are summarized in Figure 4.3. Each plot corresponds to a specific dataset utilized for evaluating the proposed training approach, using either ALFR or UAE-PUPET in the training process. The X-axis of the plot represents the size of G3 data to see how different size of G3 data affects the multi-part privacy-utility tradeoff process. The Y-axis denotes the inference accuracy obtained of both group's private and utility features. Each data point represents the highest accuracy achieved among the three aforementioned threat models, averaged over 25 experiments. Since it is a balanced classification problem, we focus solely on accuracy and the area under the curve (AUC) metric. For brevity, we exclude the AUC results as they are similar to the accuracy results. We acknowledge the existence of a trivial solution when the size of G3 is zero. However, we also conduct experiments without

employing this trivial solution when G3 is equal to zero for completeness.

In general, we observe that all three training approaches yield successful results, in the sense that they significantly reduce the private features of both G1 and G2, while maintaining a relatively high accuracy in inferring the utility features. Among the three approaches, the model-sharing approach demonstrates the best overall performance, as it achieves comparable privacy levels to the other approaches while providing higher utility. However, we do not conduct an extensive comparison to determine the absolute best approach, as this research area is still in its early stages and lacks prior studies for comparison. We encourage further engagement from researchers in this field and hope that our results can serve as a standardized benchmark for future investigations.



(a) Different attack technique used by adversaries

(b) Different evaluation techniques used by utility providers

Figure 4.4: Results showing what percentage of attack/evaluation was done by a particular type. The three points corresponding to each value on the X-axis should sum to 1.

Moreover, it is evident from Fig 4.3 that the utility is suboptimal when G3 has a very small size, specifically 0 or 10. However, as the size of G3 increases, there is an observable improvement in utility while the level of privacy remains relatively consistent. To the best of our knowledge, prior research has not explored this relationship between the size of G3 and the tradeoff between privacy and utility. Additionally, it is worth noting the selection of threat/evaluation techniques employed for different sizes of G3. Figure 4.4a demonstrates

various attack techniques utilized by adversaries to predict the private features of both groups, while Figure 4.4b shows the different techniques employed by utility providers to predict utility features of both groups. We start from G3 size equal to 10 because when G3 = 0 the first threat model can't be used and the third and second model essentially becomes the same. Both plots exhibit similarities in that, when the G3 size is only 10, both adversaries and utility providers use the second technique mentioned in Section 4.3.4 to predict the private and utility features, as it yields the highest inference accuracy. This is very much intuitive, considering the difficulty of training neural network classifiers with only 10 data points from G3 and the challenge of imputing values as described in the third technique. As the size of G3 increases, adversaries and utility providers increasingly rely on G3 data for their respective tasks. Hence, when a sufficient amount of G3 data is accessible, it becomes possible to infer the utility features and protect private features solely based on G3 data using the existing approaches without relying on the data of other groups. However, in scenarios where the size of G3 data is inadequate, the training approaches we introduce in this work are necessary.

4.5 Conclusion

In this study, we introduced a novel multi-party framework aimed at balancing privacy and utility considerations between two distinct user groups. To address this tradeoff, we proposed three training approaches: the individual approach, the data sharing approach, and the model sharing approach. Our findings indicate that when an insufficient amount of raw data is available for training a privacy mechanism, leveraging data from other groups that have already undergone privatization becomes necessary. Although our approaches demonstrate promise, there is still potential for further optimization, particularly in terms of the data sharing and model sharing approaches, which can be time-consuming due to their iterative nature.

Chapter 5

Initial Exploration of Zero-Shot Privacy Utility Tradeoffs in Tabular Data Using GPT-4

5.1 Introduction

A particular architectural progression has significantly enhanced the efficient utilization of GPUs for computational purposes: the *Transformer* architecture (Vaswani et al., 2017). Serving as the cornerstone for numerous recent innovations, the Transformer architecture has facilitated the development of various Large Language Models (LLMs). The advent of these LLMs has led to a profound transformation within the field of computer science, with unparalleled growth and advancements.

LLMs are characterized by their massive architecture, with models having billions or trillions of parameter (e.g GPT-3 model with 175 billion parameters (Brown et al., 2020), GPT-4 model rumored to have 1.7 trillion parameters). Training these models involves working with extensive datasets that tap into a wide array of information from diverse data sources. By capitalizing on insights derived from these varied sources, LLMs have a tendency

to amass a comprehensive range of knowledge, rendering them highly adaptable to a diverse array of tasks. Therefore, they have found application across various domains, including applications in sentiment analysis, text generation, code generation, multimodal tasks, and more. They have also proven effective in wide array of tasks in low-data scenarios such as zero-shot and few-shot settings (Brown et al., 2020). LLMs have not only demonstrated their effectiveness in handling unstructured data but have also exhibited remarkable performance when applied to structural tabular datasets within zero-shot and few-shot settings (Hegselmann et al., 2023).

LLMs pose a significant challenge when it comes to preserving privacy. One of the primary concerns is their inadvertent potential to reveal information about the training data (Carlini et al., 2021). This vulnerability implies that malicious individuals could potentially access the training data, which may contain sensitive information, leading to privacy breaches. Additionally, LLMs have made substantial advancements in their ability to draw inferences, potentially allowing them to deduce various personal attributes of users. Staab et al. (2023) demonstrated how large language models can deduce sensitive information from users during the inference phase. Therefore, privacy issues regarding LLMs go beyond just extracting training data; they also involve privacy violations due to LLMs' powerful inference capabilities.

Several studies such as Behnia et al. (2022); Peris et al. (2023) have proposed methods, to protect LLMs from revealing their training data. These methods focus on modifying the model's parameters using different techniques to enhance privacy. However, as of our current knowledge, there hasn't been prior research exploring whether LLMs can effectively be used to safeguard user privacy concerning their ability to make inferences. Therefore, this work shifts its focus towards investigating the potential of LLMs to leverage their inference abilities and statistical insights to maintain user privacy, especially when dealing with tabular datasets. Our main objective is to use LLMs, especially GPT-4, to sanitize datasets in a way that hinders the extraction of sensitive user information while still retaining

the ability to extract useful features. To the best of our knowledge, this study represents the first attempt to explore how LLMs can be employed to enhance the privacy of tabular datasets while preserving their usefulness.

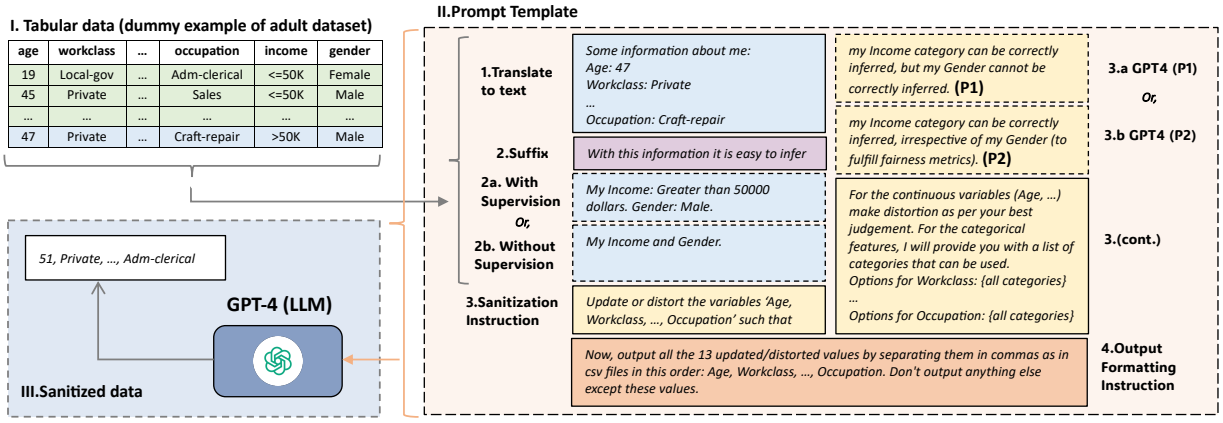


Figure 5.1: Overview of the proposed data sanitization technique using GPT-4. Our approach involves initializing the GPT-4 model with a prompt that includes the conversion of tabular data points into textual format, followed by the incorporation of sanitization instructions. Additionally, we incorporate formatting instructions into the prompt to ensure that the generated outputs are suitable for programmatic utilization. The presented prompt above pertains to the prompt designated for *Task 1*.

5.2 Related Work

Within the sphere of privacy, two fundamental categories have been established: *data privacy* and *inference privacy* (Sun et al., 2018; Sun and Tay, 2020). *Data privacy* primarily focuses on preserving the original, unaltered data in its raw state. *Inference privacy* deals with the protection of sensitive information that can be derived or deduced from the disclosed raw data, often as a consequence of correlations. In this context, *raw data* denotes the initial data collected or generated, which remains unaltered and represents the data in its fundamental state before undergoing any form of examination, modification, or inference.

In the broader discourse concerning privacy implications in LLMs, a central concern revolves around the identification of the training data. This aligns with the paradigm of *data*

privacy, wherein the actual raw data contains sensitive information, and the extraction of this data poses a substantial privacy risk. Carlini et al. (2021); Nasr et al. (2023); Li et al. (2023), demonstrate diverse methods for extracting training data from language models. To mitigate the risk of extraction of private training data, Behnia et al. (2022) introduced a method for privately fine-tuning large language models using Differential Privacy, and Peris et al. (2023) further demonstrate the fine-tuning of LLMs with enhanced privacy. In a similar vein, Wu et al. (2023) proposed a privacy neuron detector designed to identify neurons associated with private information and subsequently modify these identified privacy neurons by setting their activations to zero. Furthermore, Kassem et al. (2023) enhanced the privacy of language models through reinforcement learning, utilizing negative similarity scores to promote paraphrasing and reduce dependence on the original training data. Huang et al. (2023) utilize a straightforward sanitization step aimed at mitigating privacy risks in retrieval-based language models.

It is important to note that these methods primarily address the issue of *data privacy* and do not directly tackle the core focus of our research, which centers on *inference privacy*. Staab et al. (2023) demonstrate privacy concerns related to LLMs extend beyond mere memorization of training data and how the improved efficiency of LLMs in inference can be leveraged to infer private attributes from user texts during inference. While LLMs have predominantly focused on privacy from an unstructured data perspective (e.g text, image, etc.), Hegselmann et al. (2023) showed LLMs can be utilized for classification tasks in zero-shot and few-shot settings in tabular datasets as well. The statistical knowledge acquired by LLMs extends beyond textual data, potentially posing privacy risks in tabular datasets, as demonstrated by Staab et al. (2023), where GPT-4 is capable to infer private attributes such as birth place, race, education level, income, and gender with a high accuracy from the ACS Income Dataset (Ding et al., 2022) in a zero-shot setting.

Various adversarial optimization techniques such as Edwards and Storkey (2016); Huang et al. (2017); Madras et al. (2018); Pittaluga et al. (2018); Osia et al. (2020); Huang et al.

(2019); Chen et al. (2019); Wu et al. (2020); Morales et al. (2020); Xiao et al. (2020); Erdemir et al. (2021); Wu et al. (2021); Mandal et al. (2022), have been utilized to tackle the inherent trade-offs between safeguarding inference privacy and preserving data utility. These techniques either introduce additional noise into their generator network or latent variables or employ loss functions that achieve distortion without explicitly adding noise. They have proven their effectiveness in thwarting Machine Learning (ML) models from inferring sensitive attributes like race, gender, income, and others depending on the application. Moreover, these methods are adaptable and can be applied to different data types, including tables, images, and text (representations). The primary focus of these methodologies has revolved around ensuring inference privacy, with the aim of preventing the revelation of sensitive attributes or meeting fairness criteria within ML models. However, as pointed out in Chapter 2, a notable portion of this research has not thoroughly addressed the unique characteristics of the data and the requisite post-processing steps when dealing with tabular datasets that incorporate categorical features. The chapter emphasized that sanitized datasets generated through diverse privacy mechanisms often necessitate additional post-processing efforts to restore the data to its original format.

Our objective in this work is to explore the potential of utilizing Large Language Models (LLMs) as an alternative to adversarial optimization techniques for ensuring privacy. Given that GPT-4 has demonstrated exceptional zero-shot capabilities in accurately inferring a wide range of attributes, we aim to delve deeper into whether GPT-4 can also sanitize datasets to prevent the inference of sensitive attributes. The forthcoming section will provide a more detailed exploration of the problem that this work seeks to address using LLMs.

5.3 Problem Formulation and Methodology

Problem: Let's consider the scenario of an external service provider tasked with the responsibility of sanitizing a dataset to both protect user privacy and ensure the desired utility is

preserved. This provider has access to a raw database containing a data vector, denoted as D, which consists of various features represented as $\{d_1, d_2, d_3, \ldots, d_n\}$. The primary objective for this provider is to sanitize the dataset D in such a way that when users release their sanitized data, denoted as $\hat{D}$, it enables accurate inference of utility features, denoted as D_U, while simultaneously making it difficult to accurately infer private features, denoted as D_P. Initially, the dataset D exhibits correlations with both D_U and D_P and it's important to note that $D_U \notin D$ and $D_P \notin D$. Failing to properly sanitize and releasing D publicly could lead to the accurate inference of private features, thus compromising user privacy. In a public domain, potential attackers might utilize various pre-trained machine learning models to deduce private features from the data. Therefore, it is crucial for the sanitizing function f to be robust against multiple models to ensure effective privacy protection. In the context of this work, *privacy is construed as the pretrained model's incapacity to accurately deduce private attributes, whereas utility is characterized as the pretrained model's competence in accurately deducing utility attributes.*

Prior research has primarily focused on adversarial optimization approaches to identify an effective sanitizing function f, or other theoretical techniques involving noise addition (Sharma et al., 2021). In contrast, our approach explores the potential of using GPT-4 as a viable sanitizing function f.

Proposed Methodology: We now detail the methodology for employing GPT-4 as a sanitizing function f. In this specific context, the function f is a composite of two distinct functions, namely p and g, expressed as $f(.) = g(p(.))$. Here, p is responsible for the prompting mechanism, while g represents the pre-trained GPT-4 model. The prompting function p operates by taking the original tabular data point as its input and generating a prompt. The prompting function p encompasses several sequential steps elaborated below:

1. ***Translation to Text***: The first phase involves converting tabular data into text. Hegselmann et al. (2023) demonstrate that different techniques for this conversion produce similar outcomes, as long as all the information from the tabular data is

retained. Therefore, we adopt a straightforward technique for this conversion. The exact prompt is illustrated in Figure 5.1, where the variables {*age, workclass, ..., occupation*} represent all features in vector D.

2. ***Suffix with Supervision***: This stage involves annotating the data D with accurate private and utility labels D_P and D_U, respectively. This annotated information is subsequently incorporated into the prompt, as depicted in Figure 5.1. In this illustration, D_P corresponds to gender, while D_U corresponds to income.

3. ***Sanitization Instruction***: This portion involves the addition of prompts that instruct the model to sanitize the dataset D in order to achieve the specified objectives related to privacy and utility. We conducted an evaluation using two distinct prompts, namely $P1$ and $P2$. The precise prompts are displayed in Figure 5.1.

4. ***Output Formatting Instruction***: The final part of the prompt entails incorporating instructions for arranging the output in a specific sequence, ensuring that the outcome from the GPT-4 function g aligns with our desired format and can be seamlessly added to the database as $\hat{D}$.

Upon completion of the function p transforming the tabular data into a comprehensive prompt that encompasses all necessary guidelines for data sanitization, the resulting output of p is then supplied to the function g. The role of function g is to produce the sanitized dataset, denoted as $\hat{D}$. In formal terms, this process is represented by the equation $\hat{D} = g(p(D, D_U, D_P))$, indicating the sequential application of functions p and g to achieve the final sanitized dataset.

5.3.1 Assumptions and Threat Model

In our conceptual framework, we establish a foundational trust relationship between the data owner and the third-party service provider. This trust empowers the data owner to

confidently share raw data with the service provider, who is responsible for data sanitization. Analysts or potential attackers seeking to extract private or utility features from the data $\hat{D}$ possess an auxiliary dataset at their disposal. This auxiliary dataset enables them to initially train various machine learning models that can predict the private and utility features. After these models are trained, analysts proceed to attempt the inference of private and utility features for the users whose data is contained within $\hat{D}$.

5.3.2 Existing Sanitization Mechanisms (Baselines)

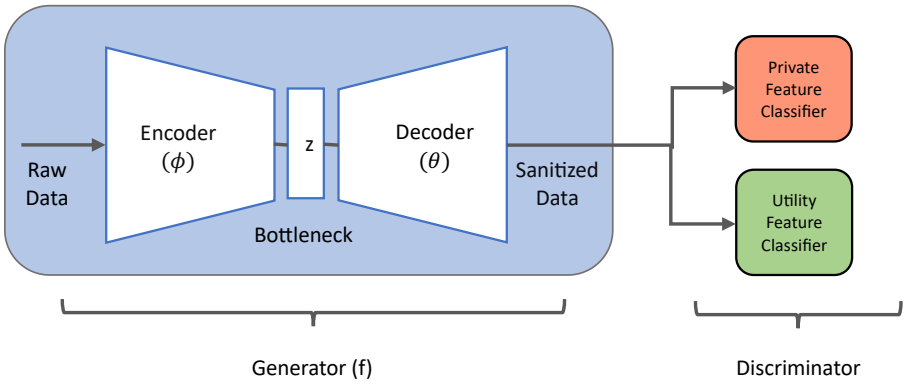


Figure 5.2: ALFR and UAE-PUPET architecture

In order to evaluate the effectiveness of GPT-4 as a privacy-preserving tool, we draw comparisons with two privacy mechanisms that employ adversarial optimization techniques: ALFR (Edwards and Storkey, 2016), and UAE-PUPET introduced in Chapter 2. These methods are characterized by a generator-discriminator framework, depicted in Figure 5.2. Within this architecture, the generator is a generative model comprising an encoder-decoder pair, while the discriminator is made up of neural network models tasked with classifying private and utility features. The discriminator's role is crucial, providing the necessary loss function that ensures the sanitized data $\hat{D}$ conceals private features while maintaining the inferability of utility features.

In the case of ALFR, we adopt the sanitization algorithm outlined in their paper and integrate it with the model architecture used in UAE-PUPET to ensure consistency. Algorithm 6 demonstrates the ALFR sanitization mechanism. In this algorithm, f, c_p, and c_u correspond to functions associated with generator, private feature classifier, utility feature classifier networks. In parallel γ, γ_p, and γ_u denote the weights of these functions, respectively. Additionally, α, λ_p, and λ_u are scalars utilized to prioritize specific loss components in the final loss function L.

Algorithm 6 ALFR training algorithm

$\gamma, \gamma_p, \gamma_u \leftarrow$ initialize weights
$U \leftarrow True$
repeat
  $\hat{D} \leftarrow f(D)$, $\hat{D}_P \leftarrow c_p(\hat{D})$, $\hat{D}_U \leftarrow c_u(\hat{D})$
  $C \leftarrow \text{MSE}(\hat{D}, D)$
  $l_p \leftarrow \text{cross\_entropy}(\hat{D}_P, D_P)$
  $l_u \leftarrow \text{cross\_entropy}(\hat{D}_U, D_U)$
  $L \leftarrow \alpha C - \lambda_p l_p + \lambda_u l_u$
  if U **then**
    Update γ_p to *maximize* the loss L
  else
    Update (γ, γ_u) to *minimize* the loss L
  end if
  $U \leftarrow$ not U
until Deadline

5.4 Experimental Setup

5.4.1 Dataset

We conduct our experiment using the UCI Adult dataset (Becker and Kohavi, 1996). Our dataset selection process involved reviewing various tabular datasets initially used for fairness studies, such as *COMPAS, Bank Marketing, Communities and Crime, Student Performance, and German Credit Dataset*, as well as others mentioned in Hegselmann et al.

(2023) like *Blood, Callhousing, Car, Diabetes, Heart, and Jungle*. The selection of the UCI Adult dataset was motivated by several considerations. Firstly, we required a dataset that encompassed a minimum of two dependent variables capable of serving as private and utility features. Secondly, our research objective involved the development of machine learning models that would not be able to accurately infer private features following data sanitization. Consequently, it was imperative to have at least two dependent variables with high initial accuracy to enable us to assess whether there was a notable alteration in model accuracy subsequent to data sanitization.

Furthermore, we sought a dataset with high zero-shot classification scores when utilized with LLMs, signifying that the language model possess statistical information pertaining to that dataset. To fulfill all these criteria, the UCI Adult dataset emerged as the only suitable choice. Although the ACS Income Dataset met these criteria as well, we ultimately opted exclusively for the UCI Adult dataset due to its similar nature and the fact that the inclusion of the ACS dataset would not significantly diversify our experiments. In order to introduce diversity into our research endeavors, we devised two distinct tasks, designated as Task 1 and Task 2.

Task 1

Task 1 focuses on the UCI Adult dataset, designating Gender as the private feature and the prediction of whether an individual's income exceeds $50,000 as the utility feature, similar to the work in Chapter 2. In this task, all features apart from the private and utility ones are classified as variables requiring sanitization, denoted as D in our problem formulation.

Task 2

In Task 2, although the features designated for sanitization (D) stay the same, there's a role reversal between the private and utility features. In this context, the private feature is now the classification task determining whether an individual's income exceeds $50,000,

and the Gender is considered the utility feature. This modification is intentionally made to demonstrate that the outcomes of our research are not influenced by the specific selection of a feature as private or another as utility and therefore, confirms the effectiveness of our proposed methodology, irrespective of which variable is designated as private or utility.

5.4.2 Performance Evaluation Metrics

To evaluate the efficacy of our sanitization mechanism, we proceed to compute the accuracy and F1 scores for an array of classifier models. Initially, we calculate these scores using the dataset in its unsanitized state. Subsequently, we assess the performance of these models after the data has undergone sanitization. Our expectation is that the accuracy and F1 scores for the utility feature should exhibit minimal decline in comparison to their original scores when the data remains unsanitized. However, for the private feature, our objective is to achieve a significant reduction in these scores.

Additionally, we employ privacy-utility tradeoff metrics introduced in Zhang et al. (2023). They define privacy leakage for attribute p as follows:

$$M_p = \frac{c_a(p) - c_r(p)}{c_n(p) - c_r(p)} \tag{5.1}$$

Similarly, utility performance of attribute u, is defined as follows:

$$M_u = \frac{c_a(u) - c_r(u)}{c_n(u) - c_r(u)} \tag{5.2}$$

We ensure that these metrics are constrained within the range of 0 to 1. In Eq. 5.1, Eq. 5.2, $c_n(.)$ represents the accuracy of a classifier in predicting either the private feature p or the utility feature u using raw, unsanitized data. $c_a(.)$ denotes the accuracy after the sanitization process, while $c_r(.)$ signifies the accuracy associated with random guessing. A

lower value of M_p indicates enhanced privacy, as it implies a closer resemblance to random guessing. Conversely, a higher value of M_u is preferable, signifying a lower drop in utility. It is noteworthy that the accuracy for random guessing in predicting Income is 0.74, while for Gender, it is 0.69 (these are unbalanced classification problems).

Since existing adversarial optimization techniques are commonly applied to achieve fair representation learning, we also investigate fairness metrics to evaluate whether our mechanism complies with various fairness standards. We consider three fairness metrics: equalized odds, equalized opportunity, and demographic parity. In brief, *equalized odds* mandates that a classifier maintains equal true positive rates and false positive rates across different sensitive groups. *Equal opportunity*, a more relaxed version of *equalized odds*, is satisfied if the classifier exhibits equal true positive rates for all sensitive groups, without specific consideration for false positive rates. Similarly, *demographic parity*, also referred to as statistical parity, necessitates that the decision rate (the rate of predicting the positive class) remains consistent across all groups, irrespective of the base rates of the outcome within those groups (Garg et al., 2020). Note that, lower scores for these metrics indicate better fairness.

5.4.3 Implementation Details

In this work, we access the GPT-4 model through OpenAI's API, specifically utilizing the *gpt-4-1106-preview* model and its chat completions functionality. We utilize the default system prompt. The model is currently subject to a daily token limit of 500,000, within which our prompt equates to roughly 710 data points. Moreover, we also use GPT-4 as a zero-shot classifier. The exact prompt used for classification tasks is shown in Figure 5.3, again with the default system prompt. The inspiration for the prompt's prefix comes from Staab et al. (2023), and as noted in Hegselmann et al. (2023), varying prompts don't significantly alter results, provided that all necessary information is clearly conveyed to the model. This classification prompt facilitates the use of about 2600 data points within the 500,000-token limit. Our reported results are based on a temperature parameter set to 0.1,

and we found that varying this parameter did not significantly alter the core outcomes of our study.

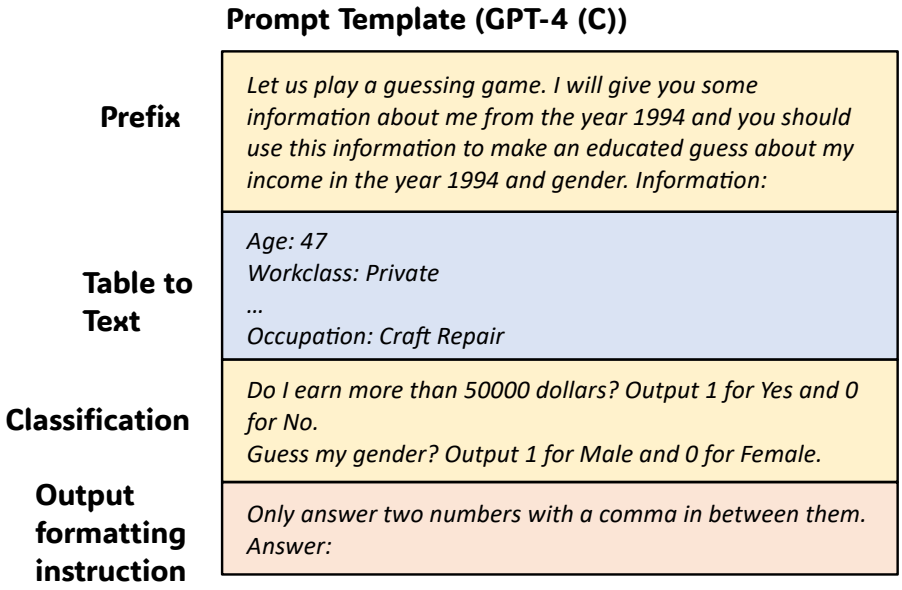


Figure 5.3: GPT-4 zero-shot classification prompt, GPT-4 (C)

5.5 Result and Discussions

In this section, we present the results obtained from our proposed methodology. We conduct a total of five independent experiments using different random seeds to calculate the mean and standard deviations of the results. Our test dataset comprises 1000 data points, a selection made with consideration to the associated costs with using the Open AI API and daily token limits. Table 5.1 shows the mean accuracy and F1 scores obtained by various models before and after the implementation of the sanitization mechanism for both tasks (standard deviation in the subscript). The *Summary* in each segment, such as No PM, ALFR, etc., indicates the highest accuracy and F1 scores achieved among the machine learning models employed, including Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), Feed Forward Neural Network (NN), and the GPT-4 Classifier (GPT-4 (C)). The highest scores

are considered as the benchmark scores in inferring the private and utility features.

Privacy	Models	Task 1				Task 2			
		Gender (Private)		Income (Utility)		Income (Private)		Gender (Utility)	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
No PM	LR	$0.83_{.002}$	$0.84_{.002}$	$0.85_{.003}$	$0.84_{.003}$	$0.85_{.003}$	$0.84_{.003}$	$0.83_{.002}$	$0.84_{.002}$
	RF	$0.84_{.002}$	$0.84_{.003}$	$0.87_{.002}$	$0.85_{.003}$	$0.87_{.002}$	$0.85_{.003}$	$0.84_{.002}$	$0.84_{.003}$
	XGB	$0.84_{.002}$	$0.84_{.003}$	$0.88_{.003}$	$0.87_{.003}$	$0.88_{.003}$	$0.87_{.003}$	$0.84_{.002}$	$0.84_{.003}$
	NN	$0.83_{.002}$	$0.81_{.002}$	$0.85_{.003}$	$0.84_{.003}$	$0.85_{.003}$	$0.84_{.003}$	$0.83_{.002}$	$0.81_{.002}$
	GPT-4 (C)	$0.80_{.002}$	$0.78_{.002}$	$0.80_{.003}$	$0.80_{.003}$	$0.80_{.003}$	$0.80_{.003}$	$0.80_{.002}$	$0.78_{.002}$
	Summary	$0.84_{.002}$	$0.84_{.002}$	$0.88_{.003}$	$0.87_{.003}$	$0.88_{.003}$	$0.87_{.003}$	$0.84_{.002}$	$0.84_{.002}$
ALFR	LR	$0.65_{.006}$	$0.64_{.007}$	$0.81_{.006}$	$0.79_{.004}$	$0.75_{.007}$	$0.70_{.005}$	$0.81_{.003}$	$0.81_{.009}$
	RF	$0.65_{.010}$	$0.63_{.009}$	$0.81_{.007}$	$0.79_{.006}$	$0.75_{.006}$	$0.68_{.008}$	$0.80_{.005}$	$0.80_{.005}$
	XGB	$0.64_{.005}$	$0.63_{.007}$	$0.81_{.008}$	$0.80_{.005}$	$0.72_{.007}$	$0.68_{.009}$	$0.80_{.006}$	$0.80_{.007}$
	NN	$0.65_{.007}$	$0.65_{.008}$	$0.81_{.010}$	$0.79_{.008}$	$0.75_{.006}$	$0.71_{.007}$	$0.79_{.004}$	$0.80_{.007}$
	GPT-4 (C)	$0.65_{.006}$	$0.62_{.007}$	$0.75_{.005}$	$0.75_{.005}$	$0.66_{.008}$	$0.67_{.012}$	$0.78_{.007}$	$0.75_{.011}$
	Summary	$\mathbf{0.65}_{.006}$	$\mathbf{0.65}_{.008}$	$0.81_{.006}$	$0.80_{.005}$	$0.75_{.006}$	$0.71_{.007}$	$0.81_{.003}$	$0.81_{.009}$
UAE-PUPET	LR	$0.67_{.009}$	$0.65_{.008}$	$0.79_{.007}$	$0.78_{.005}$	$0.73_{.011}$	$0.67_{.005}$	$0.81_{.005}$	$0.81_{.004}$
	RF	$0.67_{.006}$	$0.64_{.011}$	$0.80_{.006}$	$0.79_{.006}$	$0.74_{.008}$	$0.67_{.006}$	$0.81_{.004}$	$0.81_{.006}$
	XGB	$0.66_{.008}$	$0.64_{.010}$	$0.78_{.004}$	$0.77_{.004}$	$0.72_{.006}$	$0.69_{.009}$	$0.82_{.005}$	$0.82_{.003}$
	NN	$0.66_{.007}$	$0.64_{.008}$	$0.79_{.008}$	$0.78_{.008}$	$0.71_{.006}$	$0.68_{.009}$	$0.81_{.007}$	$0.81_{.008}$
	GPT-4 (C)	$0.65_{.008}$	$0.60_{.007}$	$0.75_{.006}$	$0.75_{.006}$	$0.60_{.009}$	$0.62_{.010}$	$0.78_{.006}$	$0.75_{.006}$
	Summary	$0.67_{.006}$	$\mathbf{0.65}_{.008}$	$0.80_{.006}$	$0.79_{.006}$	$0.74_{.008}$	$0.69_{.009}$	$\mathbf{0.82}_{.005}$	$\mathbf{0.82}_{.003}$
GPT4 (P1)	LR	$0.36_{.006}$	$0.30_{.005}$	$0.88_{.004}$	$0.88_{.004}$	$0.46_{.008}$	$0.49_{.008}$	$0.81_{.004}$	$0.82_{.007}$
	RF	$0.65_{.005}$	$0.66_{.007}$	$0.79_{.006}$	$0.74_{.005}$	$0.63_{.005}$	$0.58_{.010}$	$0.81_{.008}$	$0.81_{.006}$
	XGB	$0.35_{.008}$	$0.55_{.006}$	$0.86_{.005}$	$0.86_{.005}$	$0.67_{.006}$	$0.67_{.006}$	$0.79_{.007}$	$0.80_{.007}$
	NN	$0.47_{.008}$	$0.46_{.005}$	$0.79_{.005}$	$0.80_{.005}$	$0.47_{.011}$	$0.50_{.009}$	$0.79_{.004}$	$0.80_{.005}$
	GPT-4 (C)	$0.60_{.007}$	$0.61_{.005}$	$0.89_{.004}$	$0.89_{.005}$	$0.35_{.013}$	$0.39_{.009}$	$0.79_{.006}$	$0.78_{.007}$
	Summary	$\mathbf{0.65}_{.005}$	$0.66_{.007}$	$\mathbf{0.89}_{.004}$	$\mathbf{0.89}_{.005}$	$\mathbf{0.67}_{.006}$	$\mathbf{0.67}_{.006}$	$0.81_{.004}$	$\mathbf{0.82}_{.007}$
GPT4 (P2)	LR	$0.75_{.004}$	$0.76_{.005}$	$0.88_{.005}$	$0.88_{.005}$	$0.71_{.006}$	$0.72_{.008}$	$0.79_{.003}$	$0.80_{.005}$
	RF	$0.73_{.004}$	$0.74_{.007}$	$0.87_{.007}$	$0.87_{.006}$	$0.72_{.011}$	$0.73_{.009}$	$0.78_{.006}$	$0.79_{.007}$
	XGB	$0.74_{.008}$	$0.75_{.007}$	$0.85_{.005}$	$0.86_{.005}$	$0.74_{.011}$	$0.75_{.012}$	$0.77_{.007}$	$0.78_{.007}$
	NN	$0.73_{.006}$	$0.74_{.006}$	$0.86_{.008}$	$0.86_{.009}$	$0.69_{.010}$	$0.71_{.007}$	$0.78_{.005}$	$0.78_{.005}$
	GPT-4 (C)	$0.73_{.004}$	$0.73_{.005}$	$0.83_{.005}$	$0.83_{.006}$	$0.62_{.008}$	$0.64_{.010}$	$0.77_{.005}$	$0.77_{.006}$
	Summary	$0.75_{.004}$	$0.76_{.005}$	$0.88_{.005}$	$0.88_{.005}$	$0.74_{.011}$	$0.75_{.012}$	$0.79_{.003}$	$0.80_{.005}$

Table 5.1: Accuracy and F1 scores of multiple machine learning models evaluated both before and after the application of the sanitization mechanism. No PM denotes the scores obtained prior to the implementation of any sanitization mechanism.

For Task 1, the model predicts private feature with an accuracy of 0.84 without any sanitization mechanism, and utility feature with an accuracy of 0.88. After applying ALFR and UAE-PUPET, the accuracy for private feature significantly decreases to 0.65 and 0.67 respectively, while the accuracy for utility feature remains relatively stable at 0.81 and 0.80. Our proposed method, GPT-4 (P1) - which incorporates the P1 prompt as shown in Figure

5.1 - yields result similar to both existing sanitization mechanisms ALFR and UAE-PUPET, reducing the accuracy for inferring private features to 0.65. For utility feature, it surpasses existing methods with an accuracy of 0.89. However, our GPT-4 (P2) approach shows a higher privacy leakage, with an accuracy of 0.75 for private feature, yet it maintains better utility (accuracy of 0.88) than other adversarial techniques. For Task 2, GPT-4 (P1) shows better privacy protection and similar utility compared to existing adversarial techniques. For a comparison of F1 scores and the individual results of classifier models, kindly refer Table 5.1.

PM	**Task 1**		**Task 2**	
	M_p	M_u	M_p	M_u
ALFR	0.00	0.50	0.07	0.80
UAE-PUPET	0.00	0.42	0.00	0.87
GPT-4 (P1)	0.00	1.00	0.00	0.80
GPT-4 (P2)	0.40	1.00	0.00	0.67

Table 5.2: Privacy Leakage and Utility Performance results.

Table 5.2 showcases the performance concerning privacy leakage (M_P) and utility performance (M_U). For Task 1, our GPT-4 (P1) approach demonstrates result with no privacy leakage and no utility drop, surpassing the performance of existing techniques. Conversely, GPT-4 (P2) exhibits a higher degree of privacy leakage, although it maintains utility performance without a drop. For Task 2, our proposed techniques exhibit performance levels comparable to those of existing methods.

Numerous existing optimization strategies focus on sanitizing datasets to enhance fairness in current models. It's crucial to evaluate whether our sanitization method can similarly promote fairness. Figure 5.4 depict fairness metrics for Task 1 and Task 2, respectively. Each line in these plots represent a different sanitization mechanism, with the exception of the *No PM* line, which illustrates the fairness levels of various models before any sanitization. While the ALFR and UAE-PUPET methods improve fairness over not using any sanitization across all three fairness metrics, our GPT-4 (P1) mechanism falls short in providing

Figure 5.4: Fairness metrics for Task 1 and Task 2. The X-axis represents different ML models, and the Y-axis represents the fairness scores for those ML models, without sanitization (No PM), and after using different sanitization techniques.

comparable fairness. A closer examination of Prompt (P1) reveals it doesn't explicitly address fairness, which might explain its relatively lower fairness performance. However, it does show promise in certain scenarios, achieving the best fairness scores with some classification models (e.g., GPT-4 (C) for Equalized Odds and Equalized Opportunity, RF for Demographic Parity).

Since prompt P1 didn't include any mention of fairness, we introduce prompt P2, which specifically instructs the GPT-4 model to take fairness into account. This prompt shows notable improvement in Task 1 for equalized odds and equalized opportunity, achieving results on par with ALFR. However, for Task 2 (also for Task 1, demographic parity metric), the introduction of the prompt P2 does not enhance fairness, and in many cases, the scores worsen, except for the GPT-4 (C) classifier model, where fairness metrics are comparable to existing techniques.

This observation suggests that while our models don't uniformly perform similar to

existing techniques in fairness metrics, there is potential in leveraging LLMs with new improved LLMs in the future.

In our study, Prompt (P1) is tailored to sanitize the dataset by reducing accuracy, while Prompt P2 is designed to improve the fairness of the models. Consistent with their objectives, our results indeed show that P1 offers enhanced privacy protection, whereas P2 better addresses fairness constraints. Additionally, we experimented with another prompt that combines elements of P1 and P2, aiming to harness the strengths of both. However, this combined prompt did not yield superior results. The specific wording of this prompt (for Task 1) was: *My Income category can be correctly inferred, but my Gender cannot be correctly inferred (to fulfill fairness metrics).* This phrase was intended to balance privacy protection and fairness enhancement, but it did not lead to the anticipated improvements in the models' performance.

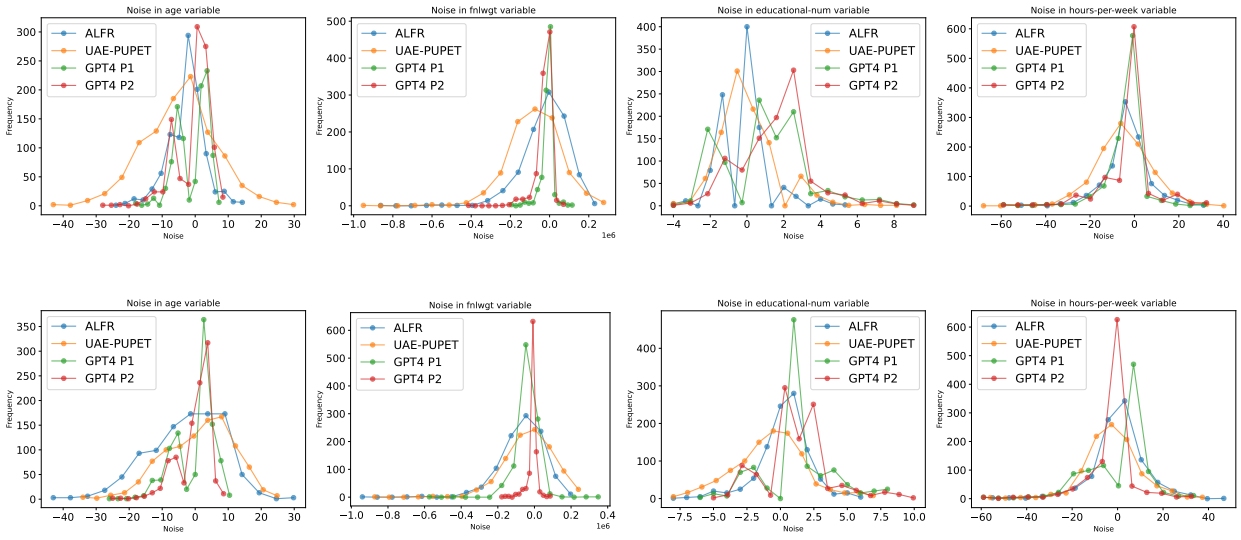
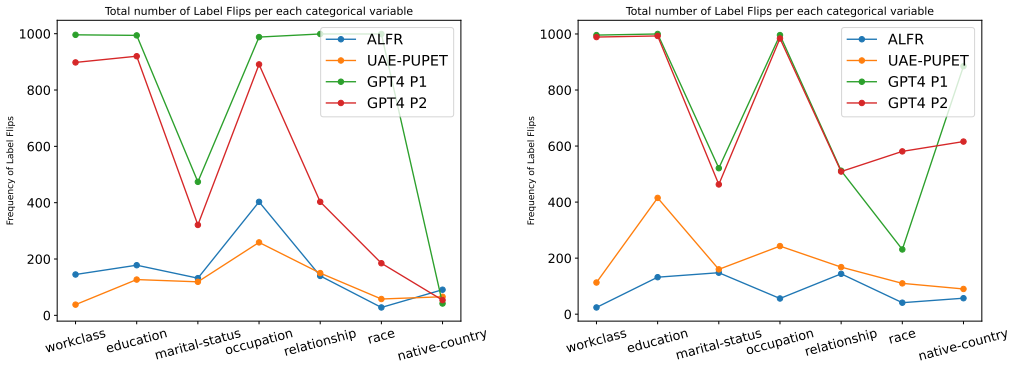


Figure 5.5: Visualization of distortion (noise) applied to continuous variables by different sanitization mechanisms for Task 1 and Task 2. Upper row shows distortion applied for Task 1 and lower row shows distortion applied for Task 2.

Furthermore, we try to gain further understanding of the distortion (noise) introduced using our proposed methodology compared to existing techniques. Therefore, we compare

the noise introduced in both existing and proposed mechanisms, to identify any potential patterns. Figure 5.5 illustrates the noise levels in continuous variables, excluding the variables *capital-gain* and *capital-loss* for the sake of brevity. The upper row of the figure depicts the noise for Task 1, while the lower row corresponds to Task 2. For continuous variables, it is observed that the noise added by our proposed mechanisms is generally less than that by the ALFR and UAE-PUPET methods, with the distributions retaining similar shapes.



(a) Task 1 (b) Task 2

Figure 5.6: Label flips for categorical variables for both tasks.

In the context of categorical variables, noise is quantified as label flips – essentially assessing whether labels change before and after the application of the sanitization mechanism. Figures 5.6a and 5.6b display the counts of label flips for each categorical variable in Task 1 and Task 2, respectively. We note that for categorical variables, the ALFR and UAE-PUPET methods result in fewer label flips compared to our GPT-4 (P1) and GPT-4 (P2) approaches, where for some variables almost all the categories are flipped.

Our study further seeks to determine whether actual private and utility labels are necessary for effective dataset sanitization. To explore this, we experimented with an unsupervised approach, as depicted in Figure 5.1, *without supervision* block, where the models were not provided with actual labels (including no range for income), to assess if they could autonomously discern the true values and sanitize the dataset accordingly. We find that including true label values, i.e., the supervised approach is important.

	Private feature			Utility feature		
	Acc	**Precision**	**F1**	**Acc**	**Precision**	**F1**
Task 1	0.68	0.73	0.69	0.78	0.78	0.78
Task 2	0.69	0.59	0.62	0.60	0.61	0.60

Table 5.3: GPT-4 (P1) sanitization without supervision.

Table 5.3 displays the highest accuracy, precision, and F1 scores achieved across the five classification models used in our study. While the outcomes for Task 1 were somewhat acceptable, the results for Task 2 clearly indicate the necessity of true labels for effective sanitization. It's also noteworthy that the highest accuracy for the utility feature in Task 1, as mentioned in the table, was achieved by only a single model. For the remaining models, there was a notable absence of preserved information for the utility feature, underscoring the challenges faced by unsupervised method in maintaining utility while protecting privacy.

5.6 Conclusion

In our study, we present a study of GPT-4 based sanitization mechanism that operates without the need for additional training, relying solely on the implementation of specifically designed prompts. Our findings reveal that this approach offers a level of privacy protection against various machine learning models that is comparable to existing techniques, despite its relative simplicity. While our method demonstrates efficacy in privacy protection, it does not consistently meet fairness constraints to the same extent as existing methods. However, certain results within our study do show promise, exhibiting comparable outcomes to those of established techniques. Given the rapid advancement in model development, we are optimistic that future iterations of such models will further enhance the capabilities of our proposed method, potentially addressing its current limitations in ensuring fairness.

Chapter 6

Concluding Remarks and Future Prospects

Social media has become an essential component of contemporary society. The plethora and diversity of information shared through social media platforms offer valuable insights for addressing significant global challenges. Concurrently, the recent advancements and expansion in machine learning present a potent means of leveraging the substantial volume of data available from social media. Nevertheless, alongside the growth in machine learning, concerns have arisen regarding the potential exploitation of these techniques by malicious entities to compromise individual privacy. In this dissertation, we delve into the realm of privacy, particularly inference privacy, examining various iterations of autoencoders and LLM to harness their potential for preserving privacy while maximizing the utility derived from the data.

In Chapter 2, we commence by delineating various privacy concerns and subsequently highlight our focus on addressing a specific privacy issue, namely inference privacy, within this dissertation. In tackling inference privacy, we introduce the Privacy and Utility Preserving End-to-End Transformation (PUPET). Our proposed approach utilizes the Uncertainty Autoencoder (UAE) as a generator model. Nonetheless, we also explore PUPET in con-

junction with different autoencoder variants such as vanilla AEs, Variational Autoencoders (VAEs), and β-VAEs, juxtaposing their performance against various techniques. Empirical analysis reveals that UAE exhibits superior performance compared to other autoencoder variants. Furthermore, within this chapter, we introduce the concept of a data type-aware technique for data sanitization, a novel approach not previously explored in existing literature.

While PUPET, particularly UAE-PUPET, was developed for single-group scenarios, wherein it effectively addresses the tradeoff between privacy and utility when all user groups share identical privacy and utility characteristics, addressing situations involving multiple user groups with differing private and utility features necessitates separate training of UAE-PUPET models for each group. However, this necessitates data annotation or reliance on auxiliary datasets for model training. In Chapter 3, we introduce a new constraint in our problem formulation, acknowledging the potential challenges in obtaining auxiliary datasets due to feasibility or cost concerns, especially when dealing with multiple user groups exhibiting diverse privacy and utility features. To resolve the privacy-utility tradeoff in such scenarios without the need for manual annotation or auxiliary datasets, we propose a data sharing mechanism. This mechanism when combined with UAE-PUPET or other exiting privacy techniques in the literature, effectively addresses the privacy-utility tradeoff for the two-user group setting, with the potential for extension to multiple user groups.

In Chapter 3, we addressed the tradeoffs between privacy and utility in scenarios involving two user groups. However, in Chapter 4, we extended the problem by introducing additional constraints, including the absence of a trusted third-party, perturbed data and labels for training privacy mechanisms, and the training of models to classify private and utility features. Collectively, Chapters 3 and 4 present novel problem formulations and corresponding solutions. It is worth noting that some of these approaches may entail increased time consumption. Therefore, future research endeavors should concentrate on refining and optimizing these methods to enhance efficiency.

In Chapter 5, our attention shifts to the capabilities of Large Language Models (LLMs), specifically GPT-4, in data sanitization. Given the recent advancements in the field of LLMs, it becomes imperative to investigate the privacy concerns and potential solutions offered by these models. Similar to Chapter 2, our focus remains on a single-group setting. This choice is deliberate as it represents the initial step and the first endeavor to utilize LLMs for inference privacy. Our objective is to assess the capabilities of LLMs in addressing a straightforward task rather than delving into the complexities associated with a two-group setting.

Throughout this dissertation, we present a range of methodologies aimed at preserving privacy while upholding the utility of data. Our findings demonstrate encouraging levels of privacy and utility assurances, as evidenced by the significant reduction in adversaries' classification accuracy when inferring private features, alongside the continued high accuracy in utility feature inference. Moreover, in our examination of inference privacy from the perspective of mitigating biases, we observe that UAE-PUPET exhibits the capacity to satisfy diverse fairness metrics.

Moving forward, the field of privacy and utility in machine learning presents several avenues for further research:

- Enhanced Adversarial Frameworks: Continued refinement of adversarial learning frameworks, particularly focusing on the development of autoencoder variants, could further improve the balance between privacy protection and data utility. Exploring alternative distribution models in the latent space may yield more flexible and efficient privacy mechanisms.

- Diversification of User Group Scenarios: Expanding upon the novel problem formulation for heterogeneous user groups, future research should aim to explore a wider array of user group dynamics. This includes considering more granular distinctions within groups and examining the impacts of varying privacy requirements on data sharing mechanisms.

- Advanced LLM Utilization: As large language models continue to advance, their potential in privacy preservation becomes increasingly significant. Future iterations of LLMs should be explored for their ability to meet fairness constraints and provide even more nuanced privacy protections without requiring extensive model training or specialized prompts.

- Cross-Disciplinary Applications: Investigating the applicability of these privacy mechanisms in various domains, such as healthcare, finance, and social media, could uncover specific challenges and opportunities in those fields. This would entail tailoring privacy-utility tradeoff solutions to meet industry-specific requirements and regulatory standards.

- Ethical and Regulatory Considerations: Further research should also delve into the ethical implications of data sanitization techniques and the evolving landscape of digital privacy regulations. Developing frameworks that not only meet technical efficacy but also align with ethical guidelines and legal standards is crucial for the widespread adoption of these technologies.

In conclusion, this dissertation lays the groundwork for innovative approaches to privacy preservation in the era of big data and machine learning. By addressing the nuances of privacy-utility tradeoffs through adversarial learning frameworks and the potential of large language models, it opens new pathways for research and application that could fundamentally alter how privacy is maintained in the digital world.

Bibliography

Acar, A., Aksu, H., Uluagac, A. S., and Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. *ACM Comput. Surv.*, 51(4).

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

Alaggan, M., Gambs, S., and Kermarrec, A.-M. (2015). Heterogeneous differential privacy.

Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., Degano, P., and Palamidessi, C. (2011). Differential privacy: on the trade-off between utility and information leakage. In *International Workshop on Formal Aspects in Security and Trust*, pages 39–54. Springer.

Asoodeh, S., Alajaji, F., and Linder, T. (2015). On maximal correlation, mutual information and data privacy. In *2015 IEEE 14th Canadian Workshop on Information Theory (CWIT)*, pages 27–31. IEEE.

Basciftci, Y. O., Wang, Y., and Ishwar, P. (2016). On privacy-utility tradeoffs for constrained data release mechanisms. In *2016 Information Theory and Applications Workshop (ITA)*, pages 1–6. IEEE.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Behnia, R., Ebrahimi, M. R., Pacheco, J., and Padmanabhan, B. (2022). Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In

2022 IEEE International Conference on Data Mining Workshops (ICDMW), pages 560–566.

Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Commun. ACM*, 13(7):422–426.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models.

Chen, J., Konrad, J., and Ishwar, P. (2018). Vgan-based image representation learning for privacy-preserving facial expression recognition.

Chen, T., Bao, H., Huang, S., Dong, L., Jiao, B., Jiang, D., Zhou, H., Li, J., and Wei, F. (2022). The-x: Privacy-preserving transformer inference with homomorphic encryption.

Chen, X., Navidi, T., Ermon, S., and Rajagopal, R. (2019). Distributed generation of privacy preserving data with user customization.

de Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2004). A tutorial on the cross-entropy method. *ANNALS OF OPERATIONS RESEARCH*, 134.

Diaz, M., Wang, H., Calmon, F. P., and Sankar, L. (2019). On the robustness of information-theoretic privacy measures and mechanisms. *IEEE Transactions on Information Theory*, 66(4):1949–1978.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2022). Retiring adult: New datasets for fair machine learning.

Domingo-Ferrer, J. and Torra, V. (2008). A critique of k-anonymity and some of its enhancements. In *2008 Third International Conference on Availability, Reliability and Security*, pages 990–993.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dwork, C. (2006). Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I., editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

Ebadi, H., Sands, D., and Schneider, G. (2015). Differential privacy: Now it's getting personal. *SIGPLAN Not.*, 50(1):69–81.

Edwards, H. and Storkey, A. (2016). Censoring representations with an adversary.

Erdemir, E., Dragotti, P. L., and Gunduz, D. (2021). Active privacy-utility trade-off against a hypothesis testing adversary.

Erdogdu, M. A. and Fawaz, N. (2015). Privacy-utility trade-off under continual observation. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1801–1805. IEEE.

Evans, D., Kolesnikov, V., and Rosulek, M. (2018). A pragmatic introduction to secure multi-party computation. *Found. Trends Priv. Secur.*, 2(2–3):70–246.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation.

Gao, S., Steeg, G. V., and Galstyan, A. (2015). Efficient estimation of mutual information for strongly dependent variables.

Garg, P., Villasenor, J., and Foggo, V. (2020). Fairness metrics: A comparative analysis.

Ghazi, B., Kreuter, B., Kumar, R., Manurangsi, P., Peng, J., Skvortsov, E., Wang, Y., and Wright, C. (2022). Multiparty reach and frequency histogram: Private, secure, and practical. *Proc. Priv. Enhancing Technol.*, 2022(1):373–395.

Ghosh, A. and Kleinberg, R. (2017). Inferential privacy guarantees for differentially private mechanisms.

Giggins, H. and Brankovic, L. (2012). Vicus: A noise addition technique for categorical data. In *Proceedings of the Tenth Australasian Data Mining Conference - Volume 134*, AusDM '12, page 139–148, AUS. Australian Computer Society, Inc.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

Grover, A. and Ermon, S. (2018). Uncertainty autoencoders: Learning compressed representations via variational information maximization.

Grover, A. and Ermon, S. (2019). Uncertainty autoencoders: Learning compressed representations via variational information maximization.

Hamm, J. and Noh, Y.-K. (2018). K-beam minimax: Efficient optimization for deep adversarial learning. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1881–1889. PMLR.

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., and Sontag, D. (2023). Tabllm: Few-shot classification of tabular data with large language models.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Hsu, H., Asoodeh, S., and Calmon, F. (2020). Obfuscation via information density estimation. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 906–917. PMLR.

Huang, C., Kairouz, P., Chen, X., Sankar, L., and Rajagopal, R. (2017). Context-aware generative adversarial privacy. *Entropy*, 19(12).

Huang, C., Kairouz, P., Chen, X., Sankar, L., and Rajagopal, R. (2019). Generative adversarial privacy.

Huang, Y., Gupta, S., Zhong, Z., Li, K., and Chen, D. (2023). Privacy implications of retrieval-based language models.

Ilanchezian, I., Vepakomma, P., Singh, A., Gupta, O., Prasanna, G. N. S., and Raskar, R. (2019). Maximal adversarial perturbations for obfuscation: Hiding certain attributes while preserving rest.

Irani, D., Webb, S., Li, K., and Pu, C. (2011). Modeling unintended personal-information leakage from multiple online social networks. *IEEE Internet Computing*, 15(3):13–19.

Kalantari, K., Sankar, L., and Kosut, O. (2017). On information-theoretic privacy with general distortion cost functions. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2865–2869. IEEE.

Kassem, A., Mahmoud, O., and Saad, S. (2023). Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In

Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4379, Singapore. Association for Computational Linguistics.

Khademi, A. and Honavar, V. (2019). Algorithmic bias in recidivism prediction: A causal perspective.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes.

Kingma, D. P. and Welling, M. (2022). Auto-encoding variational bayes.

Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., Ibrahim, M., and van der Maaten, L. (2022). Crypten: Secure multi-party computation meets machine learning.

Koufogiannis, F. and Pappas, G. J. (2016). Multi-owner multi-user privacy. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 1787–1793.

Koufogiannis, F. and Pappas, G. J. (2018). Diffusing private data over networks. *IEEE Transactions on Control of Network Systems*, 5(3):1027–1037.

LeCun, Y., Cortes, C., and Burges, C. (2010). Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Li, M., Wang, J., Wang, J., and Neel, S. (2023). Mope: Model perturbation-based privacy attacks on language models.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2017). The variational fair autoencoder.

Ma, C., Tschiatschek, S., Hernández-Lobato, J. M., Turner, R., and Zhang, C. (2020). Vaem: a deep generative model for heterogeneous mixed type data.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations.

Makhdoumi, A. and Fawaz, N. (2013a). Privacy-utility tradeoff under statistical uncertainty. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1627–1634. IEEE.

Makhdoumi, A. and Fawaz, N. (2013b). Privacy-utility tradeoff under statistical uncertainty. *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1627–1634.

Mandal, B., Amariucai, G., and Wei, S. (2022). Uncertainty-autoencoder-based privacy and utility preserving data type conscious transformation. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103.

Morales, A., Fierrez, J., Vera-Rodriguez, R., and Tolosana, R. (2020). Sensitivenets: Learning agnostic representations with application to face images.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. (2023). Scalable extraction of training data from (production) language models.

Niu, B., Chen, Y., Wang, B., Wang, Z., Li, F., and Cao, J. (2021). Adapdp: Adaptive personalized differential privacy. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10.

Osia, S. A., Taheri, A., Shamsabadi, A. S., Katevas, K., Haddadi, H., and Rabiee, H. R. (2020). Deep private-feature extraction. *IEEE Transactions on Knowledge and Data Engineering*, 32(1):54–66.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peris, C., Dupuy, C., Majmudar, J., Parikh, R., Smaili, S., Zemel, R., and Gupta, R. (2023). Privacy in the time of language models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1291–1292, New York, NY, USA. Association for Computing Machinery.

Pittaluga, F., Koppal, S. J., and Chakrabarti, A. (2018). Learning privacy preserving encodings through adversarial training.

Rajagopalan, S. R., Sankar, L., Mohajer, S., and Poor, H. V. (2011a). Smart meter privacy: A utility-privacy framework. In *2011 IEEE international conference on smart grid communications (SmartGridComm)*, pages 190–195. IEEE.

Rajagopalan, S. R., Sankar, L., Mohajer, S., and Poor, H. V. (2011b). Smart meter privacy: A utility-privacy framework. *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*.

Rassouli, B. and Gündüz, D. (2019). Optimal utility-privacy trade-off with total variation distance as a privacy measure. *IEEE Transactions on Information Forensics and Security*, 15:594–603.

Raval, N., Machanavajjhala, A., and Cox, L. P. (2017). Protecting visual secrets using adversarial nets. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1329–1332.

Raval, N., Machanavajjhala, A., and Pan, J. (2019). Olympus: Sensor privacy through utility aware obfuscation. *Proceedings on Privacy Enhancing Technologies*, 2019(1):5–25.

Rezaei, A., Xiao, C., Gao, J., Li, B., and Munir, S. (2018). Application-driven privacy-preserving data publishing with correlated attributes. *arXiv preprint arXiv:1812.10193*.

Sankar, L., Rajagopalan, S., and Poor, H. (2013a). Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852. Copyright: Copyright 2013 Elsevier B.V., All rights reserved.

Sankar, L., Rajagopalan, S. R., and Poor, H. V. (2013b). Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852.

Sarker, I. H. (2021). Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *Sn Computer Science*, 2.

Sharma, C., Mandal, B., and Amariucai, G. (2021). A practical approach to navigating the tradeoff between privacy and precise utility. In *ICC 2021 - IEEE International Conference on Communications*, pages 1–6.

Shwartz-Ziv, R. and Armon, A. (2021). Tabular data: Deep learning is not all you need.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. (2018). Learning controllable fair representations. *CoRR*, abs/1812.04218.

Staab, R., Vero, M., Balunović, M., and Vechev, M. (2023). Beyond memorization: Violating privacy via inference with large language models.

Sun, M. and Tay, W. P. (2020). On the relationship between inference and data privacy in decentralized iot networks. *IEEE Transactions on Information Forensics and Security*, 15:852–866.

Sun, M., Tay, W. P., and He, X. (2018). Toward information privacy for the internet of things: A nonparametric learning approach. *IEEE Transactions on Signal Processing*, 66(7):1734–1747.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570.

US Census Bureau (2018). Us census demographic data. https://www.kaggle.com/muonneutrino/us-census-demographic-data#acs2017_census_tract_data.csv, Last accessed on 2020-4-24.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, C. X., Song, Y., and Tay, W. P. (2021). Arbitrarily strong utility-privacy tradeoff in multi-agent systems. *IEEE Transactions on Information Forensics and Security*, 16:671–684.

Wang, H. and Calmon, F. P. (2017). An estimation-theoretic view of privacy. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 886–893. IEEE.

Wang, H., Diaz, M., Calmon, F. P., and Sankar, L. (2018). The utility cost of robust privacy guarantees. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 706–710. IEEE.

Wu, R., Zhou, J. P., Weinberger, K. Q., and Guo, C. (2022). Does label differential privacy prevent label inference attacks?

Wu, X., Li, J., Xu, M., Dong, W., Wu, S., Bian, C., and Xiong, D. (2023). Depn: Detecting and editing privacy neurons in pretrained language models.

Wu, Z., Wang, H., Wang, Z., Jin, H., and Wang, Z. (2021). Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset.

Wu, Z., Wang, Z., Wang, Z., and Jin, H. (2020). Towards privacy-preserving visual recognition via adversarial training: A pilot study.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

Xiao, T., Tsai, Y.-H., Sohn, K., Chandraker, M., and Yang, M.-H. (2020). Adversarial learning of privacy-preserving and task-oriented representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12434–12441.

Xing, Y., Song, Q., and Cheng, G. (2021). On the generalization properties of adversarial training. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 505–513. PMLR.

Zhang, L., Chen, Y., Li, A., Wang, B., Chen, Y., Li, F., Cao, J., and Niu, B. (2023). Interpreting disparate privacy-utility tradeoff in adversarial learning via attribute correlation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4690–4698.

Zhang, N. and Zhao, W. (2007). Privacy-preserving data mining systems. *Computer*, 40(4):52–58.