IACS Institute for Applied Computational Science
HARVARD SCHOOL OF ENGINEERING AND APPLIED SCIENCES

Microsoft

Microsoft: Fairness Impact of Privacy
Statement of Work

Prepared by	Kristen Grabarz, kgrabarz@g.harvard.edu Lily Ke,xiaolan_ke@g.harvard.edu Scarlett Gong, wenlin_gong@g.harvard.edu Blake Bullwinkel,jbullwinkel@fas.harvard.edu
Prepared for	Joshua Allen, joshuaa@microsoft.com
Meeting times	Mon @7pm EDT with Owen (and probably Chris),link Tue @2pm EDT for internal discussion,link Wed @7pm EDT with partner Joshua (and probably Owen),link

Background

Fairness and privacy are sometimes in conflict. Recent work has demonstrated both theoretically and experimentally that it may be difficult, or even impossible, to satisfy both of these properties at once, forcing us to make trade-offs between them. These studies have primarily focused on differentially private implementations of common machine learning algorithms, but little work has been done to investigate the trade-offs between fairness and differentially private synthetic data. SmartNoise is a Python library developed by OpenDP that has open-sourced several synthesizers that inject statistical noise into the unprotected data to generate ε-differentially private synthetic data. Our goal is to understand how changing ε (privacy loss) across various synthesizers affects our ability to achieve "fair" outcomes (according to various definitions of fairness) on common machine learning tasks such as binary classification.

Problem Statement

Our primary objective is threefold: 1) understand the tradeoff between privacy (measured by ε) and fairness (various definitions depending on the dataset) on basic tasks like binary classification. This will involve carrying out many experiments using different synthesizers (listed below) and comparing their performance in order to identify and understand conditions that lead to good or bad results. 2) Based on our findings, we hope to assess or develop pre/post-processing solutions to mitigate bias (lack of fairness) in the end results. 3) If time permits, we will also evaluate the FFPDG native-fair synthesizer from Amazon.

Resources

Datasets:

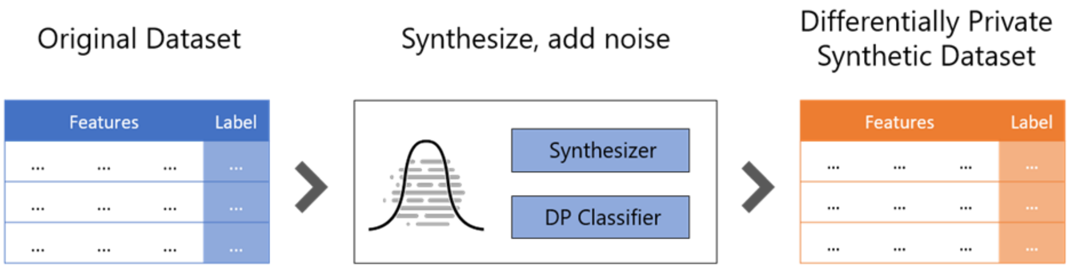
We plan on using the following datasets to evaluate synthesizer performance due to their prevalence in algorithmic fairness research. For each, we denote a machine learning task and designate potential protected attributes, e.g. groups that may be disadvantaged based on past work in the fairness literature.

- Adult Data Set
  - Predict whether income exceeds $50K/yr based on census data — binary classification. Also known as the "Census Income" dataset.
  - Protected attributes may include race (e.g., white vs. nonwhite) and gender (e.g., male vs. female).
- COMPAS
  - A landmark dataset to study algorithmic (un)fairness. This data was used to predict recidivism (whether a criminal will reoffend or not) in the USA — binary classification.
  - Protected attributes may include race (e.g., Caucasian vs. Not Caucasian) or gender (e.g., female vs. male)
- German Credit Dataset
  - This dataset predicts an individual's credit risk based on a set of personal attributes — binary classification — and also comes with a cost matrix.
  - Protected attributes may include age (e.g., young vs. old) and gender (e.g., female vs. male)

Synthesizers:

Microsoft maintains an open source library called SmartNoise that comprises mechanisms for providing differentially private results to users in order to protect the privacy of the individuals represented in the underlying dataset. One of these mechanisms relies on models that use statistical noise to generate differentially private synthetic data which preserve the statistical properties of the original dataset. With synthesizers, differentially private datasets can be analyzed while minimizing privacy risk.

While a synthetic dataset reflects many statistical features of the original data, it cannot preserve all the information contained in the original data while guaranteeing record-level privacy. This presents a fairness risk, especially for individuals who are underrepresented in the original dataset, and a broader tradeoff between fairness and privacy.



SmartNoise includes several synthesizers that generate differentially private data in different ways.

- MWEM: Combines multiplicative weights and exponential mechanism techniques; simple but effective with shorter runtime.
- DP-CTGAN: Uses CTGAN for synthesizing tabular data, applied to DPSGD. Suitable for tabular data, but can lead to expensive training times.
- PATE-GAN: Improves upon DPGAN, especially for classification; consists of two generator blocks called a student block and a teacher block on top of the existing generator block.
- DPGAN: Adds noise to the discriminator of a GAN to enforce differential privacy
- PrivBayes: Developed by DataResponsibly
- FFPDG: A synthesizer developed by Amazon. If timeallows, we are interested in comparing its performance to the core set of synthesizers in SmartNoise.

Computing Resources

For computationally expensive models such as DP-CTGAN, we plan to utilize cloud computing resources on AWS (credits provided by Chris Tanner) or Microsoft Azure.

Fairness Definitions

While our ideas will likely evolve over the course of the semester, we plan on measuring fairness using several definitions commonly found in related literature: disparate impact, equality of odds, equality of opportunity, and statistical parity.

Project Timeline

Date	Goals
Sept 22nd (ignite talk)	● Complete brief literature review ● Decide on definitions of fairness (informed by lit review) ● Download and prepare all datasets ● pip install a couple synthesizers (MWEM, DP-CTGAN) and generate some fake data
Oct 6th (milestone 1 wk)	● Prepare fairness evaluation pipelines for binary classification on each dataset (input = data, output = fairness metrics) ● Prepare fairness evaluation pipelines for hypothesis test on each dataset
Nov 3rd (milestone 2 wk)	● Calculate and analyze all results (across tasks, datasets, and synthesizers) in order to characterize the trade-off between DP and fairness ● Understand the conditions that lead to more or less fair outcomes ● Start thinking about pre/post-processing methods that could mitigate the bias we observe
Nov 17th (milestone 3 wk)	● Recommend or develop pre/post-processing methods that promote fair outcomes in the end results
Dec 15th (final deliverables)	● Any other extensions we have time for ● Finish writing final paper ● Prepare and practice presentation

Deliverables

Deliverable 1	Working data pipeline that takes in dataset and calculates relevant fairness metrics. This will be used to generate baseline fairness performance on the original datasets.
Deliverable 2	Results of running fairness evaluation pipeline on synthetic data produced by two synthesizers.
Deliverable 3	Evaluation of pre/post-processing tactics which will address: 1. Details and implementation of processing method 2. Its impact on fairness metrics using above pipeline
Deliverable 4	Research report containing a detailed analysis and comparison of different synthesizers and their fairness implications.