USENIX Security '25 Artifact Appendix: Disparate Privacy Vulnerability: Targeted Attribute Inference Attacks and Defenses

Ehsanul Kabir
Pennsylvania State University

Lucas Craig
Pennsylvania State University

Shagufta Mehnaz
Pennsylvania State University

A Artifact Appendix

A.1 Abstract

As machine learning (ML) technologies become increasingly prevalent in privacy-sensitive domains such as healthcare and finance, they raise critical concerns about potential privacy leakage risks. This paper investigates one such threat: attribute inference attacks, where adversaries exploit public, non-sensitive attributes to infer private, sensitive information. While these attacks may perform poorly on an entire dataset, they can achieve high accuracy for records belonging to specific vulnerable groups, a phenomenon termed disparate vulnerability. This work introduces a novel disparity inference attack to identify high-risk groups and two targeted variations of attribute inference attacks that significantly outperform untargeted versions. Additionally, we propose an innovative defense mechanism, the Balanced Correlation Defense (BCorr), which effectively mitigates these risks while preserving model utility. For artifact evaluation, our goals are to ensure reproducibility of the experiments, validate the implementation of the proposed attacks and defenses, and enable further exploration of disparate vulnerability in ML systems. The provided artifacts include datasets, a modular codebase, and detailed Jupyter notebooks for reproducing key results, such as the relationship between correlation and attack performance, the efficacy of targeted attacks, and the mitigation impact of BCorr. These artifacts aim to support open science practices, ensuring accessibility, transparency, and reusability for the research community.

A.2 Description & Requirements

This section outlines the experimental setup, including security considerations, access instructions, software dependencies, hardware requirements, and benchmark datasets for reproducibility.

A.2.1 Security, privacy, and ethical concerns

There are no known risks to machine security, data privacy, or ethical concerns associated with executing this artifact. The provided code does not involve destructive steps, nor does it disable any security mechanisms.

A.2.2 How to access

The artifact is available at https://zenodo.org/records/14732956. Please refer to 'Version v2' of the artifact for the most up-to-date version.

A.2.3 Hardware dependencies

None

A.2.4 Software dependencies

The artifact requires a Linux, macOS, or Windows operating system with Python 3.9.16 or later. A virtual environment can be set up using Conda or Python's `venv`. All necessary dependencies are listed in `requirements.txt` and can be installed using `pip install -r requirements.txt`. Running Jupyter notebooks requires installing `notebook` via `pip install notebook`. No proprietary software is required, and all dependencies are open-source and can be installed using standard package managers.

A.2.5 Benchmarks

The experiments in this artifact rely on the following datasets: `census19.csv`, which contains the Census-19 dataset, and `Adult_35222.csv` and `Adult_10000.csv`, which contain the Adult dataset partitions for training and testing, respectively. Additionally, the Texas-100X dataset is required. To ensure reproducibility, we have attached a preprocessed version of the dataset with the filename `texas_100_cleaned.csv`.

A.3 Set-up

A.3.1 Installation

To install and set up the artifact, first create a virtual environment using Conda or Python's `venv` and activate it. Then, install all required dependencies using `pip install -r requirements.txt`. To run Jupyter notebooks, install `notebook` via `pip install notebook`. Create a directory `<PATH_TO_MODELS>` to store models after training.

A.3.2 Basic Test

To verify the setup, open and run all cells in the Jupyter notebook `basic_test.ipynb`. This notebook loads and preprocesses the datasets for four different experiment scenarios. A successful run will produce the correlation values for different groups in each scenario in the final cell.

A.4 Evaluation workflow

This section outlines the steps required to evaluate the artifact, validate its functionality, and reproduce the key results presented in the paper. It consists of two parts: Major Claims, which enumerate the core findings supported by the experiments, and Experiments, which describe the operational steps needed to test these claims. Each experiment is linked to a corresponding claim and includes details on execution, estimated compute time, and expected outcomes.

A.4.1 Major Claims

(C1): The correlation between the sensitive attribute and the output is a significant factor in the vulnerability to attribute inference attacks. Specifically, datasets with high correlation are more vulnerable than datasets with low correlation. This is proven by the experiment (E1) described in section 4.1 of the main paper whose results are illustrated in Figure 1.

(C2): Correlation influences disparate vulnerability among groups i.e. groups with high correlation are more vulnerable to attribute inference attacks than groups with low correlation. Additionally, the correlation level influences the confidence score distribution and can be estimated using angular difference. This is proven by the experiment (E2) described in section 4.2 of the main paper whose results are illustrated in Figure 2 and Figure 3.

(C3): Imputation attack performance degrades if the auxiliary dataset size is small or the marginal prior of the auxiliary dataset is different from the target data. Additionally, practical imputation attack (ImpP) performance is poorer than CSMIA/LOMIA in high correlation groups. This is proven by the experiment (E3) described in section 6.2 of the main paper whose results are illustrated in Figure 4.

(C4): Disparity inference attack can rank groups in terms of vulnerability more efficiently than the baseline attack using auxiliary dataset. This is proven by the experiment (E4) described in section 6.3 of the main paper whose results are reported in Table 1.

(C5): Single Attribute-based and Nested Attribute-based targeted attribute inference attacks can achieve higher attack success rate than their untargeted counterpart. Specifically, the attack performance gradually increases with lower attack budget (κ). This is proven by the experiment (E5) described in section 6.4 of the main paper whose results are reported in Table 2 and Table 3.

(C6): Our proposed defense BCorr can effectively mitigate disparate vulnerability without reducing model utility or introducing fairness concerns. Specifically, ASRD is negligible after applying BCorr and there is a minimal change in model performance and fairness metrics such as Equalized Odds Difference (EOD) and Demographic Parity Difference (DPD) do not increase. This is proven by the experiment (E6) described in section 7.2 of the main paper whose results are reported in Table 5.

A.4.2 Experiments

(E1): [3 compute-hour]: This experiment evaluates the impact of correlation between sensitive attributes and model outputs on the success of various attribute inference attacks. Using the Census19 and Texas-100X datasets, we generate 19 training sets per dataset, each with varying levels of correlation. We then train target models and perform CSMIA, LOMIA, Imputation, and Neuron Importance attacks. The results are expected to indicate that CSMIA and LOMIA exhibit a strong positive relationship with correlation, achieving higher attack success rates as correlation increases.
Preparation: No additional preparation required.
Execution: Run all cells in the Jupyter notebook `correlation_vs_attack_performance.ipynb`
Results: The last two cells of the notebook is expected to generate plots similar to Figure 1.

(E2): [2 compute-hour]: This experiment examines the impact of correlation on disparate vulnerability by analyzing how attack performance varies across different groups within the Census-19 dataset. Specifically, we vary correlation levels in Male and Female groups across nine scenarios, observing that groups with higher correlation experience greater attack success rates. Additionally, we introduce the concept of angular difference by analyzing confidence score distributions using hexagonal binning. The result is expected to reveal that attack vulnerability is closely tied to correlation, with distinct trajectory patterns in confidence scores across different groups with varying correlation.
Preparation: No additional preparation required.
Execution: Run all cells in the Jupyter notebook `angular_difference_by_sex.ipynb`
Results: The last two subsections of the notebook are expected to generate plots similar to Figure 2 and Figure 3 respectively. Specifically, in each scenario, the group with the higher correlation should show higher attack success rate than the group with the lower correlation. Additionally, angular difference is plotted for groups with varying correlation (-0.4, -0.5, -0.6) and a trend of increase in angular difference should be observed with higher correlation value.

(E3): [1 compute-hour]: The experiment evaluates the impact of distributional drift in auxiliary datasets on the performance of imputation attacks. Two types of drift are analyzed: dataset-level distributional drift, where the marginal prior and dataset size is varied to assess how deviations from the original training data affect attack performance, and group-level distributional drift, where correlations within occupation groups differ from the overall dataset correlation. The result is expected to demonstrate that imputation attacks perform well only when the auxiliary dataset closely matches the original data distribution. In cases where marginal prior deviates significantly or group-level correlations shift, imputation attacks perform worse than CSMIA and LOMIA, particularly in highly vulnerable groups.
Preparation: No additional preparation required.
Execution: Run all cells in the Jupyter notebook `imputation_vs_ai_aux_size_and_distrib_diff.ipynb`
Results: The subsection 'Plot Results' and the last subsection of the notebook are expected to generate plots similar to Figure 4(a) and Figure 4(b) respectively. Specifically, the generated table should show lower attack performance as dataset size decreases or marginal prior decreases. The plot should report lower attack performance for ImpP than CSMIA/LOMIA for groups 'Adm-clerical', 'Prof-specialty', and 'Tech-support'.

(E4): [1 compute-hour]: This experiment evaluates the effectiveness of the proposed disparity inference attack, which ranks groups based on their vulnerability to attribute inference attacks. To assess the ranking quality, the experiment uses Kendall's Tau and Spearman's Rank Correlation, comparing the disparity inference attack against a baseline ranking method that relies on an auxiliary dataset.
Preparation: No additional preparation required.
Execution: Run all cells in the Jupyter notebook `disparity_inference.ipynb`
Results: The last two cells of the notebook are expected to generate results similar to Table 1. Specifically, the ranking quality (Spearman's R/Kendall's Tau) of disparity inference attack should be significantly better (closer to ± 1) than baseline.

(E5): [1 compute-hour]: This experiment evaluates two proposed variations of targeted attribute inference attacks—single attribute-based and nested attribute-based—using the Census19, Texas-100X, and Adult datasets. Groups are selected based on specific attributes, with the nested approach refining selection through multiple attributes. The result is expected to show that targeted attacks consistently outperform untargeted approaches, with refined selection improving inference accuracy.
Preparation: No additional preparation required.
Execution: Run all cells in the Jupyter notebook `targeted_attribute_inference.ipynb`
Results: The subsections 'Single Attribute-based Targeted Attacks' and 'Nested Attribute-based Targeted Attacks' of the notebook are expected to generate results similar to Table 2 and Table 3 respectively. Specifically, there should be a trend in attack performance increase with lower values of κ and higher depth. Note that the untargeted attack performance in Census19 and Texas-100X may be slightly different from the ones reported in the paper. This may be caused by the randomization involved in the sampling process for these dataset specific scenarios. However, the claim in our paper that targeted attacks achieve increasingly better performance with smaller target sets is expected to uphold.

(E6): [5 compute-hour]: This experiment evaluates the effectiveness of BCorr in mitigating disparate vulnerability across different groups using the Census-19 and Texas-100X datasets. The defense is tested on both binary attributes (`SEX` and `SEX_CODE` respectively) and multi-valued attributes (`ST` and `PAT_STATUS` respectively). For comparison, a Fairness Constraint-based defense (FC) is also applied as a baseline. The evaluation measures attack success rate disparity (ASRD), group fairness (EOD, DPD), and model accuracy (MA) to assess the impact of BCorr on both security and fairness. The result is expected to show that BCorr significantly reduces disparities across groups while maintaining model utility and preserving fairness.
Preparation: No additional preparation required.
Execution: Run all cells in the Jupyter notebook `balancing_corr_defense.ipynb`
Results: The last cell of the notebook is expected to generate results similar to Table 5. ASRD with BCorr should be significantly smaller than without BCorr across all scenarios. MA with BCorr is expected to be similar to MA without BCorr. EOD and DPD with BCorr should not be higher than EOD and DPD without BCorr.

A.5 Version

Based on the LaTeX template for Artifact Evaluation V20231005. Submission, reviewing and badging methodology followed for the evaluation of this artifact can be found at https://secartifacts.github.io/usenixsec2025/.