Diffusion Models for the Evaluation of Interference-Cognizant Estimators

Freek

January 19, 2025

Click here for the latest version

Abstract

I demonstrate using state-of-the-art generative artificial intelligence models called diffusion models to simulate credible economic network data. I also show how the method generalizes to non-network economic data. In the method, (i) treatment assignment can be controlled, (ii) true treatment effects are known, and (iii) there is limited room for researcher discretion in shaping the simulated data output. I apply the proposed methodology to a dataset of student social networks in US middle schools. Leveraging the correspondingly simulated data, I compare the performance of popular estimators accounting for spillovers and provide recommendations for their use.

1. Introduction

In the aftermath of the COVID-19 pandemic, there has been a growing use of and interest in casual estimators that account for spillovers between units (Google Scholar 2024). These estimators are not only applicable to problems in vaccine distribution but in general apply to empirical research featuring interference between units. Example settings include those of housing voucher allocation (Hudgens and Halloran 2008) and social experiments (Aronow and Samii 2017).

When running such empirical research, it is of interest to know about the practical robustness of the different estimators that account for spillovers. Researchers conventionally test the practical performance with Monte Carlo studies. The credibility of these studies is limited because of researcher discretion in choosing the design (Athey et al. 2019), and because these studies are unlikely to be informative in real-world settings (Advani, Kitagawa, and Słoczyński 2018). Therefore, making practical performance comparisons between different estimators accounting for spillovers would be more powerful when the simulated data used is more realistic, and leaves fewer degrees of freedom in the simulation stage.

Athey et al. (2019) use methods in generative artificial intelligence that can be used to generate economic data that is more realistic than that from Monte Carlo simulations. In particular, the authors investigate the performance of Wasserstein Generative Adversarial Networks (WGANs) (Goodfellow et al. 2014; Arjovsky, Chintala, and Bottou 2017) in generating realistic economic data. The training data in this analysis are the results of a program evaluation experiment. In this setting, the Stable Unit Treatment Values (SUTVA) assumption is implied by the type of estimators that Athey et al. compare to each other, probably motivated by test subjects representing an insignificant part of the overall labor market (Lalonde 1986; Dehejia and Wahba 1999, 2002).

In this paper, I simulate data with interference between units using methods in generative artificial intelligence, and compare the performance of estimators that account for interference between units. I choose to use diffusion models to simulate the data over GANs because diffusion models (Ho, Jain, and Abbeel 2020) display superior generative capabilities over GAN models (Yang et al. 2022), and because diffusion models do not feature the inherent instability of GAN models during training and sampling (Cao et al. 2022). These issues would become more pervasive as the size of interference-featuring network training data grows quadratically in the number of units.

I train a diffusion model on social network data of an anti-conflict experiment within 56 schools, in which influential students are randomly assigned to be eligible to participate in an anti-conflict training program (Paluck, Shepherd, and Aronow 2016). This setting is suitable for this research because (i) we can expect spillovers between students in a school, (ii) treatment is assigned at random within a well-defined group of students, meaning we can expect a fitted diffusion model to give unbiased predictions of outcomes under counterfactual treatment assignments using the same mechanism, (iii) there is variation in the proportion of students treated within each contained unit (a school), which will allow the diffusion model to model the relationship between treatment proportion intensity and spillover structures, and (iv) knowing how the performance of estimators accounting for spillovers compares in the setting of school experiments will be informative to empirical research in education, and may be relevant to empirical research in other settings of social networks.

(Conjectured) I find that the relative performance of different estimators on the generated data depends on the sample size, the demographic breakdown of counterfactual schools, and the counterfactual assignment structure of treatment. I provide guidance on picking an estimator that accounts for spillovers in three frequent empirical situations.

Sections 2-6, respectively, present 2 the mathematical background of diffusion models, 3 the social network data, 4 an overview of the simulated data, 5 the comparison of estimators, and 6 the discussion including a comparison between WGAN models and diffusion models with a note on the recovery of privacy-sensitive data.

2. Diffusion Models

2.1. Forward and Backward Processes

A diffusion model consists of two main components: the diffuser and the denoiser. The diffuser progressively adds noise to the training data points. The denoiser learns to remove the noise at each iteration to recover the original data points. After training a denoiser, one can draw a vector of random noise and have the denoiser iteratively dust away some noise to eventually recover an artificially generated data point that is reminiscent of the training data.

Denote a data point by $\mathbf{x}_0$, with $\mathbf{x}_0 \in \mathbb{R}^d$. In our setting, $\mathbf{x}_0$ is a simple weighted undirected network of units with unit characteristics. With that, $d = \binom{n}{2} + nm$ if there are n units, there are m unit-level characteristic variables, and weights are a real number. In the empirical setting of this paper, weighted edges will be friendship strengths between students, and characteristic variables will be demographic variables such as gender and age.

The diffuser iteratively adds noise to training data points according to the rule

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\varepsilon}_t, \tag{1}$$

where $\boldsymbol{\varepsilon}_t$ is a standard Gaussian, β_t is a fixed scalar with $\beta_t \in (0,1)$, and $t \in \{1, 2, \ldots, T\}$. This rule thus "diffuses" data as t grows. The rule from equation (1) defines the discrete time Markov process with the transition density

$$q\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right) = \mathcal{N}\left(\mathbf{x}_t;\ \sqrt{1-\beta_t}\mathbf{x}_{t-1},\ \beta_t\mathbf{I}\right), \tag{2}$$

where $\mathbf{I}$ is the identity matrix. One can show that we can generalize to the transition density

$$q\left(\mathbf{x}_t|\mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_t;\ \sqrt{\bar{\alpha}_t}\mathbf{x}_0,\ (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{3}$$

where $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{\tau=1}^{t}\alpha_\tau$. One can also show that

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\bar{\epsilon}_t\right) + \sqrt{\frac{(1-\alpha_t)\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_{t-1}}}\mathbf{z}, \tag{4}$$

where $\bar{\epsilon}_t = \prod_{\tau=1}^{t}\boldsymbol{\varepsilon}_\tau$, and $\mathbf{z}$ is a standard Gaussian.

2.2. Algorithms

We can train our diffusion model by (i) selecting records from the training data at random, (ii) adding noise to those data points according to randomly sampled data points from $\mathrm{Unif}\{1, 2, \ldots, T\}$ using equation (3), and (iii) training a neural network to predict $\bar{\epsilon}_t$ that recovers $\mathbf{x}_0$, under predetermined α_t for all $t \in \{1, 2, \ldots, T\}$.

These steps can be more formally expressed in the following algorithm.

Algorithm 1 Training the diffusion model

repeat
  $\mathbf{x}_0 \sim X$ ▷ *Draw a data point from the training set*
  $t \sim \mathrm{Uniform}\{1, 2, \ldots, T\}$ ▷ *Pick a perturbation progression level*
  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ *Pick an error term as to draw* $\mathbf{x}_t$

  Take a step away from the loss' gradient:
  $\nabla_\theta \|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon},\ t\right)\|$ ▷ *Predict* $\boldsymbol{\varepsilon}$ *according to equation (3)*
until converged

After training, we can generate artificial observations using equation (4) as expressed by the following algorithm.

Algorithm 2 Sampling from the diffusion model

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ *Draw a completely diffused "data point"*
2: **for** $t = T, T-1, \ldots, 1$ **do** ▷ *Progressively denoise the diffused data point*
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
4: $\quad \mathbf{x}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\right)\boldsymbol{\varepsilon}_\theta + \sqrt{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_{t-1}}}\mathbf{z}$ ▷ *Apply equation (4)*
5: **return** $\mathbf{x}_0$

3. Data

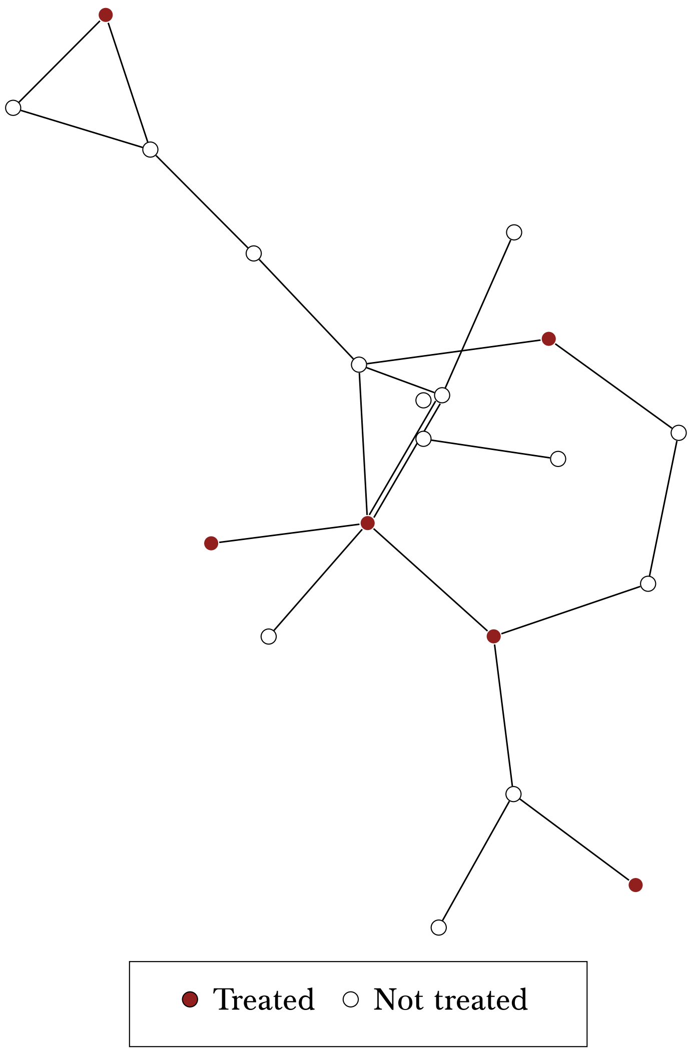
The data I use in this study is from a randomized experiment of student-specific anti-conflict training with 56 public middle schools is New Jersey (Paluck, Shepherd, and Aronow 2016). In each of the 28 treated schools, a number of "influential" students are randomly selected to participate in the experiments. The researchers have all students in the 56 schools fill out a survey before and after the intervention. The surveys ask each student to list which other students they interact with, and who their two best friends are. The surveys also collect demographic data. With the resulting survey responses, one can construct a social network of the students in each of the 56 schools.

4. Generating Student Network Data

With the data as described in section 3, I start with training a diffusion model on small random samples of 20 students within school-grades. As (i) the relationship between any two students is taken as symmetric, and (ii) I retain 3 demographic variables per student, this process involves implementing a diffusion model following Algorithm 1 and 2 to generate data points with a dimensionality of $\binom{n}{2} + nm = \binom{20}{2} + 20 \cdot 3 = 250$[1]. I will also generate network data for samples of 50 and 100 later. For now, the following acts as a proof of concept.

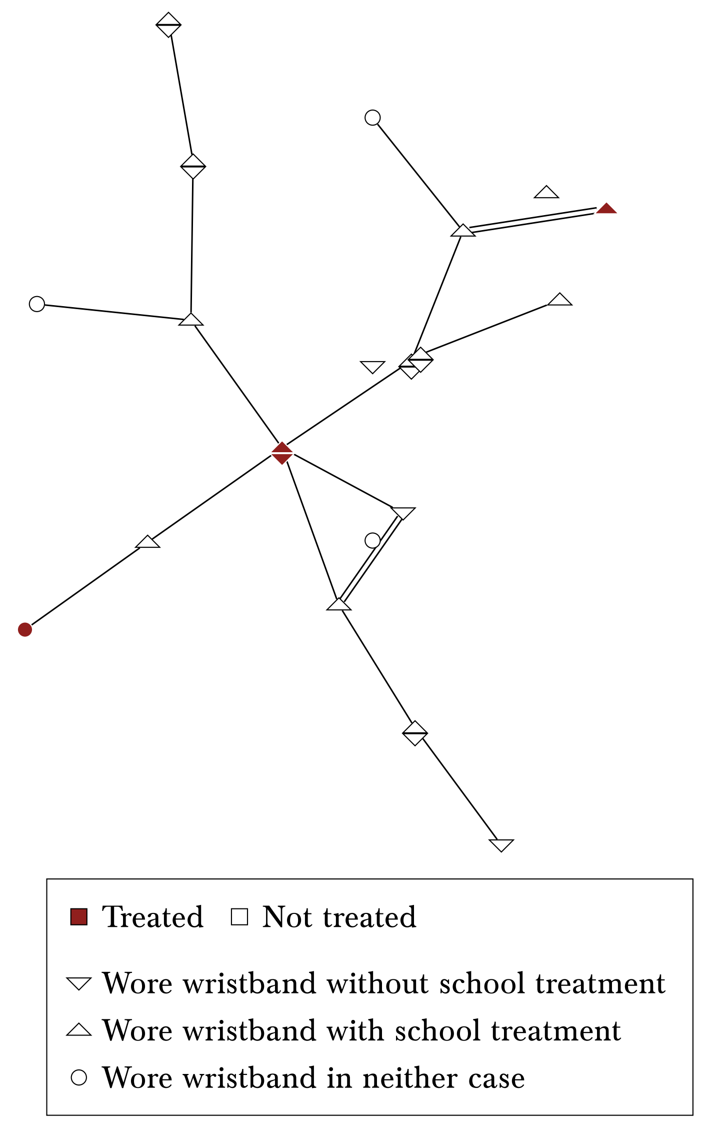
[1] I am using the equation here which I cover in the introduction.

Figure 1: Example of Training Sample



Notes: This figure shows the relationships between and treatment of 20 students in a sample of the training data. Vertices in this graph represent students. When there is an edge between two vertices in this graph, then one of the associated students claimed they spent time with the other in the last few weeks. When there is a double edge, then both students said the other is their best friend or without loss of generality student 1 said student 2 is one of their best friends and student 2 said they spent time with student 1 in the last few weeks. A student is treated when they were selected for a program to encourage a dismissive public stance against conflict at school.

Figure 2: Example of Generated Sample



Notes: This figure shows the relationships between and treatment of students as part of a generated sample. Vertices in this graph represent students. When there is an edge between two vertices in this graph, then one of the associated students is predicted to have spent time with the other in the last few weeks. When there is a double edge, then both students are predicted to have said the other is their best friend or without loss of generality it is predicted that student 1 said student 2 is one of their best friends and student 2 said they spent time with student 1 in the last few weeks. Treatment is the selection into a program to encourage a dismissive public stance against conflict at school. Wearing a wristband to communicate support to the anti-conflict program is an outcome variable shown in the graph. “Without school treatment” refers to the predicted outcomes in the counterfactual in which no students are treated. “With school treatment” refers to treatment of students in the sample as shown in the graph.

5. Comparison Estimators Accounting for Interference

In this section, I will compare the empirical coverage of the true treatment effects under varying treatment assignment schedules of these estimators applied to a diffusion model that is trained like described in section 2.2. The estimators I pick are consistent with those in the analysis in (Aronow and Samii 2017):

1. The Thompson estimator
2. The Hajek estimator
3. Weighted Least Squares
4. Ordinary Least Squares
5. The difference in sample means (DSM) estimator
6. An exposure-based estimator

6. Discussion

In this section of this paper, I will discuss the benefits of evaluating estimators with artificial data generated with DDPM or WGAN models, and its limitations.

I will also compare the generative performance of WGAN models and diffusion models. See section A of the appendix for the current comparisons.

References

Advani, Arun, Toru Kitagawa, and Tymon Słoczyński (2018). *Mostly Harmless Simulations? Using Monte Carlo Studies for Estimator Selection*. eprint: arXiv:1809.09527.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). *Wasserstein GAN*. eprint: arXiv:1701.07875.

Aronow, Peter M. and Cyrus Samii (2017). "Estimating Average Causal Effects Under General Interference, With Application to a Social Network Experiment". In: *The Annals of Applied Statistics* 11.4, pp. 1912–1947. ISSN: 19326157. URL: http://www.jstor.org/stable/26362172 (visited on 07/01/2024).

Athey, Susan et al. (2019). *Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations*. eprint: arXiv:1909.02210.

Cao, Hanqun et al. (2022). *A Survey on Generative Diffusion Model*. eprint: arXiv:2209.02646.

Dehejia, Rajeev H and Sadek Wahba (2002). "Propensity score-matching methods for nonexperimental causal studies". In: *Review of Economics and statistics* 84.1, pp. 151–161.

— (1999). "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs". In: *Journal of the American Statistical Association* 94.448, pp. 1053–1062. ISSN: 01621459. URL: http://www.jstor.org/stable/2669919 (visited on 07/11/2024).

Goodfellow, Ian J. et al. (2014). *Generative Adversarial Networks*. eprint: arXiv:1406.2661.

Google Scholar (2024). *Toward causal inference with interference*. URL: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=VvTOFXkAAAAJ&citation_for_view=VvTOFXkAAAAJ:YsMSGLbcyi4C.

Higham, Catherine F., Desmond J. Higham, and Peter Grindrod (2023). *Diffusion Models for Generative Artificial Intelligence: An Introduction for Applied Mathematicians*. eprint: arXiv:2312.14977.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). *Denoising Diffusion Probabilistic Models*. eprint: arXiv:2006.11239.

Hudgens, Michael G and M Elizabeth Halloran (2008). "Toward causal inference with interference". In: *Journal of the American Statistical Association* 103.482, pp. 832–842.

LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". In: *The American Economic Review* 76.4, pp. 604–620. ISSN: 00028282. URL: http://www.jstor.org/stable/1806062 (visited on 07/11/2024).

Munro, Evan, Xu Kuang, and Stefan Wager (2021). *Treatment Effects in Market Equilibrium*. eprint: arXiv:2109.11647.

Paluck, Elizabeth Levy, Hana Shepherd, and Peter M. Aronow (2016). "Changing climates of conflict: A social network experiment in 56 schools". In: *Proceedings of the National Academy of Sciences* 113.3, pp. 566–571. DOI: 10.1073/pnas.1514483113. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1514483113. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1514483113.

Yang, Ling et al. (2022). *Diffusion Models: A Comprehensive Survey of Methods and Applications*. eprint: arXiv:2209.00796.

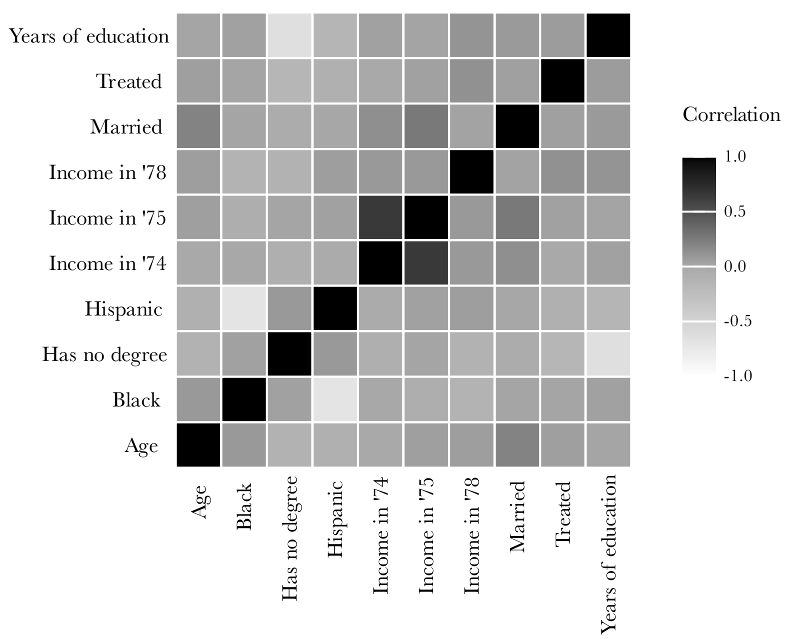
Appendix

Table of Contents

A. Comparison Generative Ability LDW Data

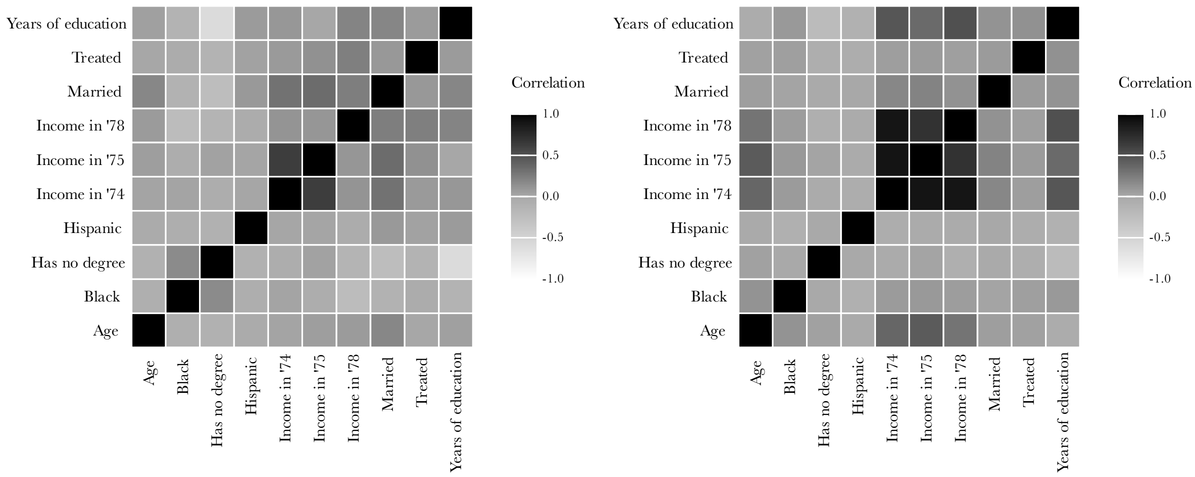
A.1. Correlation Structures

Figure 3: Correlation Matrix Real Data



Notes: This figure displays the correlations between different demographic variables in the LDW dataset.

Figure 4: Correlation Matrices Simulated Data



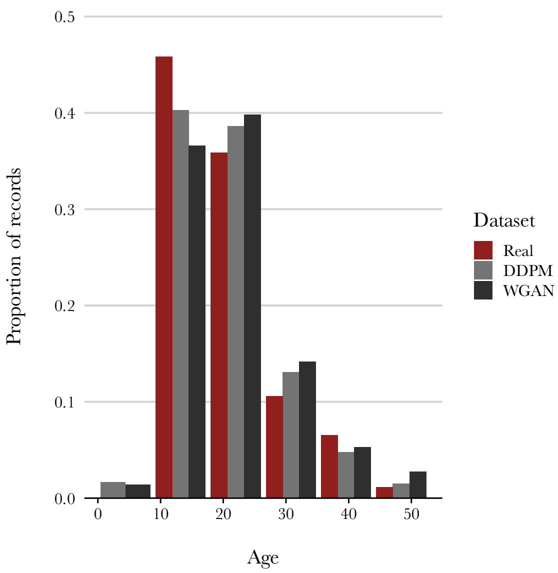
(a) DDPM model

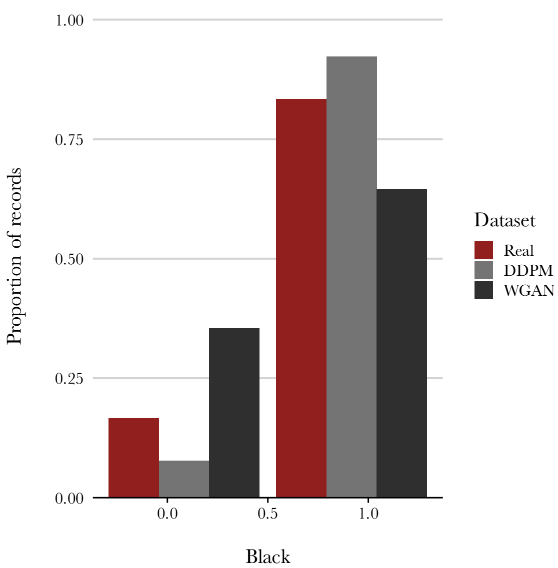
(b) WGAN model

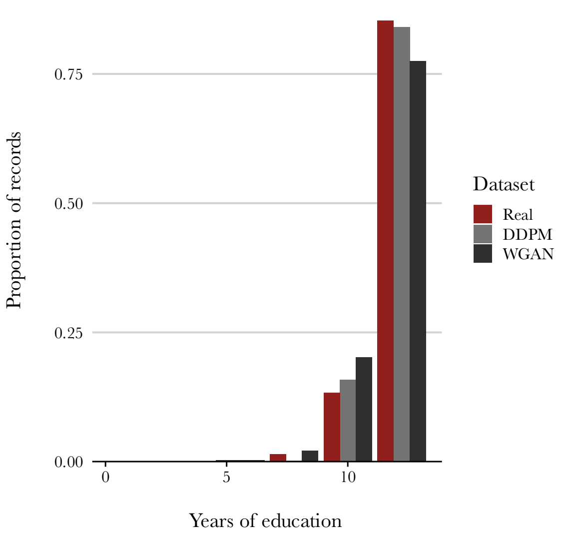
Notes: This figure displays the correlations between the different demographic variables of artificially simulated data. The simulated data comes from a Denoiser Diffusion Probabilistic Model (DDPM) and a Wasserstein Generative Adversarial Network (WGAN). It appears that the DDPM does a better job at mimicking the correlation structure of the original data than the WGAN model.

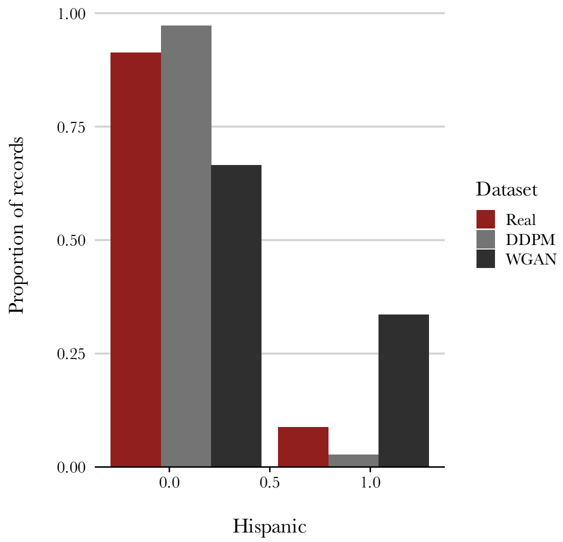
A.2. Marginal Distributions

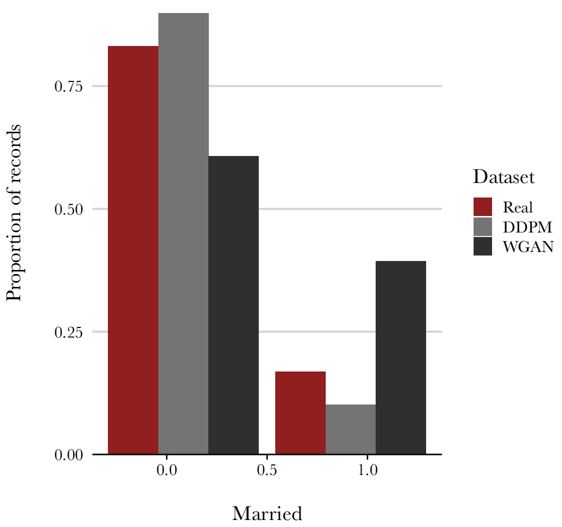
In the below figures, I present the marginal distributions and (in one plot) the conditional marginal distributions of the demographic variables in the Lalonde-Deheijia Wahba (LDW) data set.

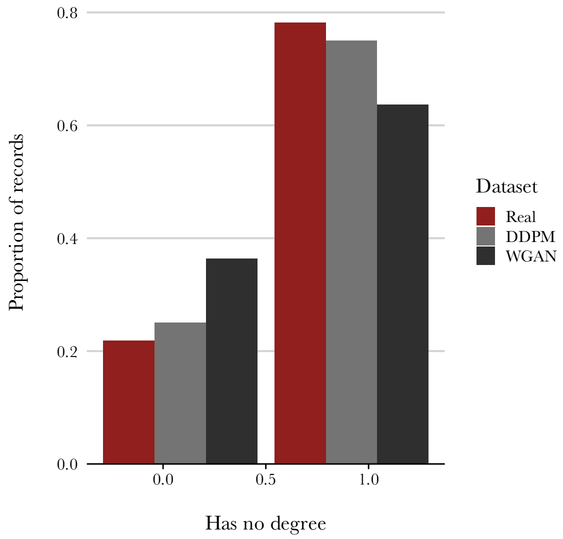


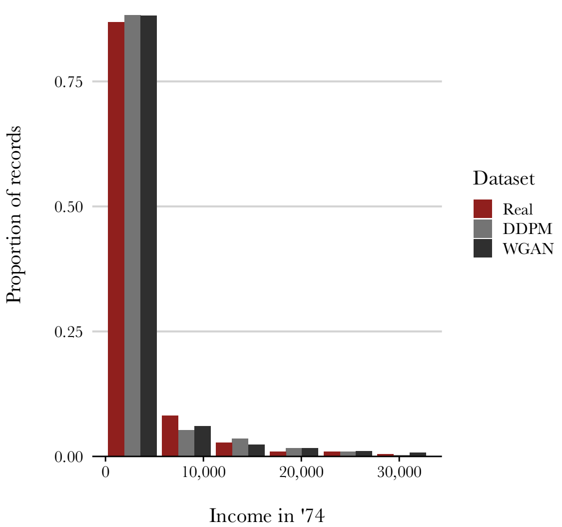


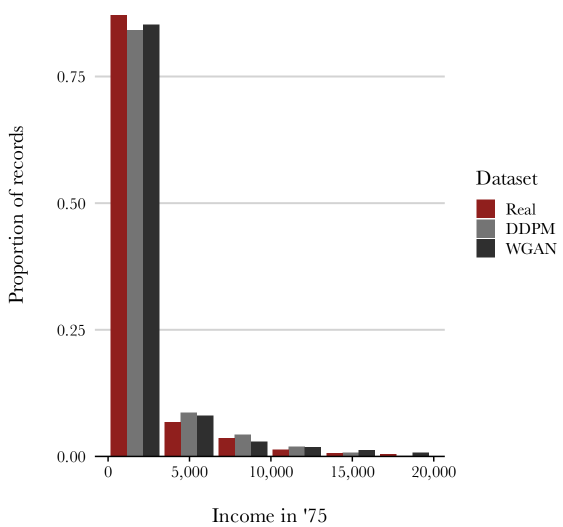

Proportion of records
0.00
0.25
0.50
0.75
0
5
10
Years of education
Dataset
Real
DDPM
WGAN

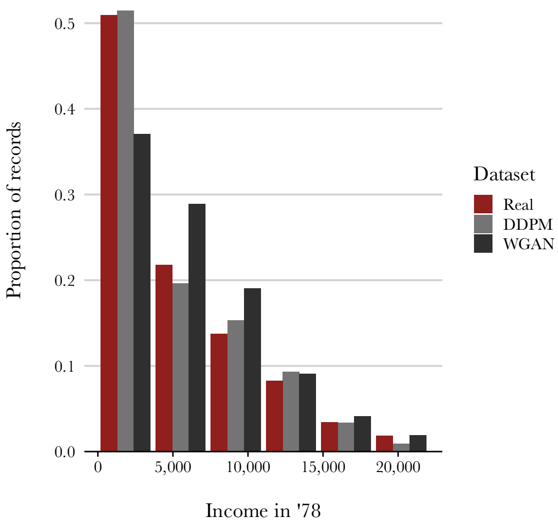

Proportion of records
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0
Hispanic
Dataset
Real
DDPM
WGAN

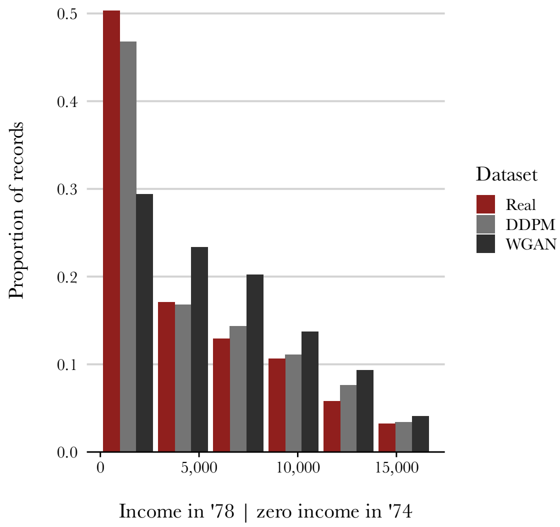

Proportion of records
0.00
0.25
0.50
0.75
0.0
0.5
1.0
Married
Dataset
Real
DDPM
WGAN


Proportion of records
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
Has no degree
Dataset
Real
DDPM
WGAN


Proportion of records
0.00
0.25
0.50
0.75
0
10,000
20,000
30,000
Income in '74
Dataset
Real
DDPM
WGAN


Proportion of records
0.00
0.25
0.50
0.75
0
5,000
10,000
15,000
20,000
Income in '75
Dataset
Real
DDPM
WGAN


Proportion of records
0.5
0.4
0.3
0.2
0.1
0.0
0
5,000
10,000
15,000
20,000
Income in '78
Dataset
Real
DDPM
WGAN


Proportion of records
0.5
0.4
0.3
0.2
0.1
0.0
0
5,000
10,000
15,000
Income in '78 | zero income in '74
Dataset
Real
DDPM
WGAN