An Empirical Evaluation of Differentially Private Generative Adversarial Networks for Social Science

Working Paper

Christian Arnold
Cardiff University
CF10 3AT, UK
arnoldc6@cardiff.ac.uk

Marcel Neunhoeffer
University of Mannheim
A5 6, D-68159 Mannheim, Germany
mneunhoe@mail.uni-mannheim.de

Sebastian Sternberg
University of Mannheim
A5 6, D-68159 Mannheim, Germany
ssternbe@mail.uni-mannheim.de

September 2, 2019

Abstract

Political scientists pervasively use data that contains sensitive information – e.g. micro-level data about individuals. However, researchers face a dilemma: while data has to be publicly available to make research reproducible, information about individuals needs to be protected. Synthetic copies of original data can address this concern, because ideally they contain all relevant statistical characteristics without disclosing private information. But generating synthetic data that captures–eventually undiscovered–statistical relationships is challenging. Moreover, it so far remains unsolved to fully control the amount of information disclosed during this process. To that end differentially private generative adversarial networks (DP-GANs) have been proposed in the (computer science) literature. We experimentally evaluate the trade-off between data utility and privacy protection in a simulation study by looking at evaluation metrics that are important for social scientists, specifically in terms of regression coefficients, marginal distributions and correlation structures. Our findings suggest that on average, higher levels of provided privacy negatively affects the synthetic data quality. We hope to encourage inter-disciplinary work between computer scientists and social scientists to develop more powerful DP-GANs in the future.

***K*eywords** GAN · Machine Learning · Synthetic Data · Differential Privacy

1 Introduction

We live in a society that is collecting data at an unprecedented level – and very often this data is sensitive. Scientific studies often use confidential data. Government agencies hold data about citizens that are equally sensitive. And companies —- think Facebook, Google or Twitter —- are amassing data about consumer (online) behavior. Ideally, this data could be shared: to make studies replicable, or to externalise services for example. But of course, important privacy concerns do not allow to distribute data freely. Original data simply cannot be shared if high levels of data protection need to be guaranteed.

To address this privacy challenge, one solution is to de-identify data, for example by stripping ID variables from a data set. However, several prominent studies show that this is not enough (e.g. Narayanan & Shmatikov, 2008)[1].

This poses a formidable challenge to social scientists. Take for example an elite survey on electoral candidates. Its respondents are individuals of high public interest. Answers to researchers' questions will contain very sensitive data, for example on their campaign budget, age, gender, social media activity or other personal characteristics that could affect the candidates' electoral success. Deleting all personal information from the data set is not enough to preserve a

[1] A famous real-world example of such a linkage attack is Sweeney (1997). She re-identified records in de-identified health insurance records by matching publicly available voting data on a combination of birth date, sex, and the ZIP code of patients and was able to single out the health records of the Governor of Massachusetts. Another examples includes the successful re-identification of user data in the Netflix Prize (Narayanan & Shmatikov, 2008)

candidate's privacy. Certain data – for example the electoral district and the candidate's election results – will have to remain in the data, simply to be able to reproduce the results in the original study. Given that it is not possible to rule out re-identification even in the light of anonymization, researchers cannot simply share this data publicly. The privacy of the candidates must be protected.

Instead of excluding certain information from the original data, it is much more promising to release completely synthetic data, instead. The idea behind synthetic data is that it is ideally no longer necessary to share original data. Instead, data users receive a synthetic copy of the original data that they can analyse and even further share themselves freely: Synthetic data promises to conserve the statistical utility of the original data – it allows to draw similar inferences as with the original data. At the same time synthetic data guards the privacy of individuals in the original data set. This idea dates back to Rubin (1993) and Little (1993) who first advanced synthetic data methods in the spirit of multiple imputation. After being formulated as a proper framework (Raghunathan, 2003), a series of papers elaborated it further (Abowd & Lane, 2004; Abowd & Woodcock, 2004; Reiter & Raghunathan, 2007; Drechsler & Reiter, 2010; Kinney et al., 2010, 2011). Different statistical models to generate the synthetic data exist, including modern machine learning techniques (Reiter, 2005; Caiola & Reiter, 2010; Drechsler & Reiter, 2011).

However, even synthetic data has their limitations and they are not immune to privacy attacks. For example, outliers are hard to mask when generating synthetic data. Imagine a particularly young member of parliament from a small party. It is statistically very challenging to "mask" the characteristics in a synthetic data set. Moreover, attackers might use adversarial machine learning techniques for re-identification. It is possible to abuse query APIs to train adversarial models that could potentially reconstruct the original training data (Bellovin et al., 2018). Data system designers ultimately face the challenge that their system has to be "future proof". Just because they did not think of a particular attack does not mean it cannot exist. This holds for threats that are already technically present today. But it also includes eventual future attacks that might have unthinkable capacities – be it new algorithms, computational power or available side-data that can be matched to the synthetic data.

Over the last decade, a more rigorous disclosure protection standard – differential privacy (DP) – became increasingly popular in academia, business and government agencies alike (Abowd & Schmutte, 2019). DP originated in cryptography and provides mathematically proven privacy guarantees (Dwork et al., 2006). The general idea of DP is the requirement of an outcome of a randomized data analysis procedure (a simple statistic such as a mean, or the output of an algorithm such as a data synthesizer) not to change much when this outcome is calculated from two neighboring data sets that differ by only one record (e.g. one individual). In other words, DP guarantees that the difference between any two adjacent data sets does not disclose useful information about any individual observation. This privacy framework has recently been combined with data synthesizers, allowing to generate synthetic data with a complete control over privacy. One class of algorithms, Generative Adversarial Networks (GANs), have turned out to be particularly promising. While GANs are particularly well suited to generate synthetic data at high utility, while the DP concepts provides strict guarantees for protecting privacy.

In this paper, we systematically evaluate DP-GANs from a social science perspective. DP-GANs have been shown to work in specific cases such a the privacy preserving generation of image data. However, they have not been extensively tested regarding their capacity in generating synthetic copies of social data – the core analytical tools of social scientists. We experimentally evaluate the trade-off between data utility and privacy protection in a simulation study by looking at evaluation metrics that are important for social scientists. In particular, we assess how varying levels of privacy protection and varying dimensionality of the data affect the usability of the generated synthetic data in terms of regression coefficients, marginal distributions and correlation structures.

Our paper addresses two audiences. On the one hand, we show Social Scientists how to generate synthetic micro-data with privacy guarantees. We introduce to the concept of differential privacy and Generative Adversarial Nets as a promising data synthesizer and illustrate promises and pitfalls. On the other hand, we also turn to Computer Scientists. In highlighting the limitations of generating differentially private social science data, we intend to point to avenues for future research. Our contribution therefore follows the ultimate goal to spark a conversation and exchange between Social Scientists and Computer Scientists about privacy preserving synthetic micro data.

2 Differential Privacy and Synthetic Data Using Generative Adverserial Networks

We now briefly introduce Differential Privacy and Generative Adversarial Nets. We also discuss current applications of DP-GANs in the context of computer science applications.

2.1 Differential Privacy

Differential Privacy is a mathematical concept that offers strong privacy guarantees (Dwork, 2006; Dwork & Roth, 2013a; Nissim et al., 2017). With roots in cryptography, DP becomes increasingly popular due to its strict mathematical formulations and provable guarantees.

The primary goal of DP is to maximize the accuracy of queries from a database while minimizing the potential for privacy leakage. In essence, DP states that any operation on a data set should reveal (almost) no information that is specific to an individual within that data set (Page et al., 2018, 15). *Specific* refers to information that could not be inferred about the individual from statistical facts about other individuals. In other words, "differential privacy mathematically guarantees that anyone seeing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis." (Nissim et al., 2017, 2).

To illustrate, consider an example taken from Page et al. (2018). First, an analysis is run on a complete data set with all individuals. A second subsequent analysis then uses data that differs from the first one in only one single detail: it does not include the data on one particular individual. Differential Privacy states that the outcome distribution will be approximately the same for each scenario, and therefore very little information specific to the certain individual can be learned from the outcome of the analysis.

DP is typically achieved by adding uncertainty into the outcome of an analysis – often done in form of adding noise. A simple example might help again to grasp this idea[2]. Imagine a scenario in which someone asks several people the question: "Do you like ice cream?". The answer can either be yes or no. Now, each individual answer could be modified with a coin that is tossed prior to answering. If it is head, the interviewee gives an honest answer. If it is tails, the interviewee provides a "random" answer, determined by another coin toss. Statistically, it is still possible to deduce the overall average probability of all people who like ice cream. But each individual can now protect her privacy – after all, she might not have answered truthfully. This is exactly the basic idea of DP: because of the introduction of randomness, any individual may now be in the data base or not (Bellovin et al., 2018, 20). Social Scientists note that this is pretty much how randomized response for sensitive survey items works (see Warner, 1965).

More formally, (ϵ, δ)-DP is defined by Dwork & Roth (2013a) as:

A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ϵ, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $||x - y||_1 \leq 1$[3]:

$$Pr[\mathcal{M}(x) \in \mathcal{S}] \leq e^{\epsilon} Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta.$$

For $\delta = 0$, differential privacy is called "pure DP" or also "ϵ-DP".

ϵ is the so called privacy loss parameter which represents the strength of the privacy guarantee. ϵ is a non-negative numerical value, and can be used as a means of quantifying privacy risks. In essence, it quantifies how much information can be learned about any individual included in x, in a worst case scenario of an attacker with arbitrary side knowledge and arbitrary computational power. For small values of ϵ the potential privacy loss for any observation in x is small. Larger values of ϵ mean that the potential privacy loss is higher, with $\epsilon = \infty$ meaning that all the information about any individual can be learnt (i.e. by publishing the original data set).

For our application we consider a popular relaxation of ϵ-DP, with $\delta > 0$, so called (ϵ, δ)-DP. In the words of Dwork & Roth (2013a): "$(\epsilon, 0)$-differential privacy ensures that, for every run of the mechanism $\mathcal{M}(x)$, the output observed is (almost) equally likely to be observed on every neighboring database, simultaneously. In contrast (ϵ, δ)-differential privacy says that for every pair of neighboring databases x, y, it is extremely unlikely that, *ex-post facto* the observed value $\mathcal{M}(x)$ will be much more or much less likely to be generated when the database is x than when the database is y" (18).

What makes DP so strong is that due to its principled, mathematical definition it does not depend on attacker capabilities – an attacker can have arbitrary side knowledge and even arbitrary computational power. This is an important advantage of DP over other existing risk measures in the Statistical Disclosure Limitation (SDL) literature, which usually makes assumptions about the attacker. Most importantly this means that differential privacy is future proof. No matter what future data release will happen or how powerful computers become, the privacy guarantee will hold.

[2] This illustrative example is adopted from Bellovin et al. (2018), and goes back to Dwork & Roth (2013b).

[3] This means that x and y are two adjacent data sets that are differing only by one row.

2.2 Generative Adversarial Nets as a Means to Generate Complex Synthetic Data

GANs have become the go-to synthesizer for synthetic data generation. GANs belong to a class of deep learning algorithms that are able to achieve high-quality synthetic data due to their ability to learn underlying data distributions. They are thus excellent in generating high quality "fake" samples that are hard to differentiate from real ones – even if the GAN generates complex data structures such as images or video footage.[4]

The basic idea of a GAN is surprisingly intuitive. At its core, a GAN is a minimax game with two competing actors. A discriminator (D) tries to tell real from synthetic samples and a generator (G) intends to produce realistic synthetic samples from random noise. We use the same illustrative example as Goodfellow et al. (2014) to make GANs (and the adjustments later on) more accessible: "The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles."

In a GAN, the team of counterfeiters (Generator G) is a neural network which is trained to produce realistic synthetic data examples from random noise. The police (discriminator D) is a neural network with the goal to distinguish fake data from real data. The generator network trains to fool the discriminator network. It uses the feedback of the discriminator to generate increasingly realistic "fake" data that should eventually be indistinguishable from the original ones. At the same time, the discriminator is constantly adapting to the more and more improving generating abilities of the generator. Thus, the "threshold" where the discriminator is fooled increases along with the generator's capability to convincingly generate data similar to the original data. This goes on until equilibrium is reached[5].

Formally, this two-player minimax game can be written as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}\Big[\log D(x)\Big] + \mathbb{E}_{z \sim p_z(z)}\Big[\log(1 - D(G(z)))\Big] \tag{1}$$

where $p_{data}(x)$ is the distribution of the real data, X is a sample from $p_{data}(x)$. The generator network $G(z)$ takes as input z from $p(z)$, where z is a random sample from a probability distribution $p(z)$[6]. Passing the noise z through G then generates a sample of synthetic data feeding the discriminator $D(x)$. The discriminator receives as input a set of labeled data, either real (x) from $p_{data}(x)$ or generated $(G(z))$, and is trained to distinguish between real data and synthetic data[7]. D is trained to maximize the probability of assigning the correct label to training examples and samples from $G(z)$. G is trained to minimize $log(1 - D(G(z)))$. Ultimately, the goal of the discriminator is to maximize function V, whereas the goal of the generator is to minimize it.

The equilibrium point for the GANs is that the G should model the real data and D should output the probability of 0.5 as the generated data is same as the real data – that is, it is not sure if the new data coming from the generator is real or fake with equal probability.[8]

2.3 Applications of DP-GANs

It is straightforward to implement differential privacy within the GAN framework to generate privacy preserving synthetic data. First attempts have been made early on (e.g. Abadi et al., 2016). But the development of these so-called DP-GANs has recently seen a sharp increase. What ultimately makes GANs differentially private is the injection of the right amount of noise into the training process of the discriminator. Abadi et al. (2016) lay an important foundation for later applications to GANs. They clip gradients and add Gaussian noise at each update step, and then calculate the overall privacy loss (ϵ, δ) using so-called moments accounting.

A number of studies follows the framework of Abadi et al. (2016). Beaulieu-Jones et al. (2017) generate differentially private data from a medical trial. Xie et al. (2018) also produce differentially private data: they too ensure differential privacy by combining noise addition and weight clipping. Triastcyn & Boi Faltings (2018) use differentially private GANs to explicitly hide sensitive data. They enforce DP on the penultimate layer by clipping its ℓ^2 norm and adding Gaussian noise. Pairwise comparisons between every possible pair of adjacent data sets allows to evaluate the privacy

[4] Notably, GANs may rely on task-specific neural networks, e.g. CNN or RNN.

[5] Interestingly, a GAN is therefore a dynamic system where the optimisation process is not seeking a minimum, but an equilibrium instead. This is in stark contrast to standard deep learning systems, where the entire loss landscape is static.

[6] Usually GANs are set up to either sample from uniform or Gaussian distributions.

[7] This is a standard binary classification problem, and thus the standard binary cross-entropy loss with a sigmoid function at the end can be used.

[8] Note the connection to the measure of general utility presented by Snoke et al. (2018). The explicit goal of a GAN is to maximize general utility, and therefore a natural way to generate fully synthetic data.

parameter ϵ empirically. Finally, Xu et al. (2019) apply a combination of adding noise and gradient pruning, and use the Wasserstein distance as an approximation of the distance between probability distributions.

These existing DP-GANs are often benchmarked on large data sets using image data, for instance images of handwritten digits (MNIST)[9] or celebrity face images (CelebA)[10].

Using image data for the evaluation of synthetic data has a great advantage: Even though images are a very complex distribution to learn, the human eye is excellent in recognising the quality of the generated synthetic data. For example, it is possible to generate artificial data at different levels of ϵ and then plot the resulting images. The trade-off between trade-off between data utility and privacy protection can be evaluated with the bare eye.

However, social scientists are often interested in other properties of the data: what happens to other metrics such as regression coefficients, marginal distributions and correlation structures? In particular, are regression coefficients still useful when models are calculated using synthetic data? What happens to the marginal distributions and correlation structures when noise is added to the training process? Because these questions have not been addressed before, in the following sections we use a simulation study to give first answers to these questions and evaluate the trade-off between data utility and privacy protection from a social scientists perspective.

3 An Empirical Evaluation of DP-GANs

We now empirically evaluate DP-GANs from a Social Scientist's perspective, focusing in particular on the trade-off between data utility and privacy protection. We turn to a simulation study and evaluate the usability of the synthetic data generated by a DP-GAN with respect to a measure of general utility for synthetic data (Snoke et al., 2018). Furthermore, we correlation structures and regression coefficients in resulting synthetic data at various levels of ϵ.

3.1 Data for the Simulation Study

To empirically evaluate the DP-GAN, we set up an artificial and typical Social Sciences data generating process (DGP). More specifically, we rely on a DGP proposed by Montgomery & Olivella (2018). We generate $N = 1,000$ observations of 40 explanatory variables and one dependent variable. The explanatory 40 variables include "symmetric and asymmetric variables, continuous and categorical variables, and correlated and independent variables" (Montgomery & Olivella, 2018, 736) and are generated as follows:

$$\begin{gathered}
x_{1i} \sim Gamma(8,2); \\
x_{2i} \sim Gamma(10,1); \\
\begin{bmatrix} x_3 & x_4 & x_5 \end{bmatrix}_i' \sim MVN(\begin{bmatrix} 2 & 3 & 6 \end{bmatrix}, \begin{bmatrix} 1.5 & 0.5 & 3.3 \end{bmatrix}' I_3); \\
\begin{bmatrix} x_6 & x_7 & x_8 \end{bmatrix}_i' \sim Multinom(\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}, n = 1) \\
\begin{bmatrix} x_9 & x_{10} \end{bmatrix}_i' \sim MVN\left(\begin{bmatrix} -0.3 & 2 \end{bmatrix}, \begin{bmatrix} 1.5 & 0.685 \\ 0.685 & 5.5 \end{bmatrix}\right) \\
\begin{bmatrix} x_{11} & \dots & x_{40} \end{bmatrix}_i' \sim MVN(\mu, I_{30})
\end{gathered}$$

With μ being a sample of 30 integers sampling with replacement from the integers 2 to 10. Our experimental simulation data set thus contains a mix of continuous and discrete variables, as it is often the case with social science data. We also generate an outcome variable Y – Social Scientists like to run regression – as follows:

$$y = x_1 + x_2 + x_3 + x_{10} + \epsilon \tag{2}$$

where $\epsilon \sim N(0, 1)$.

The regression coefficients of a simple linear regression are expected to be $\beta_0 = 0$, and all other $\beta_{1,\dots,4}$ are expected to be 1. Knowing the true data generating process allows us to evaluate how off the coefficients of a regression are if the regression is calculated using the synthetic data set instead of the original data set.

[9] MNIST: http://yann.lecun.com/exdb/mnist/.

[10] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

3.2 DP-GAN as Data Synthesizer

We train our DP-GAN on a dataset with $1,000$ observations. The GAN runs for 500 epochs using a mini-batch size of $\frac{N}{100}$ and a learning rate $\alpha = 0.001$. The generator consists of three layers: the first hidden layer takes 100 data points from the noise vector as input to each of its 128 nodes. The second hidden layer has 64 nodes and the output layer uses these 64 observations and uses one node per feature to generate the output of the same dimensions as the original data. The output of the hidden layers is transformed by a ReLU activation function[11]. The discriminator has an almost symmetric architecture: the first layer takes the number of features (the dimensions of the original data set) as input to each of its 128 nodes. The second hidden layer has 64 nodes. The final output layer reads these 64 observations and generates the output with one node per feature: a prediction of the probability that the data comes from the real data. While the hidden layers employs ReLU activation functions, the output layer uses a sigmoid activation function[12] to output probabilities between 0 and 1. Finally, we define a fixed parameter $\delta = \frac{1}{N}$ to define the the first part of the (ϵ, δ) Differential Privacy. Clipping the ℓ^2-norm of the gradients to some finite value adds appropriate noise for training the discriminator under differential privacy.[13]. Following Beaulieu-Jones et al. (2017) we use a clipping value of 0.0001, which means that any gradient value < -0.0001 or > 0.0001 will be set to -0.0001 or 0.0001 respectively. We deploy the neural network in tensorflow and make use of its differential privacy module[14]. The GAN is optimized with DP-Adam. The tensorflow privacy module implements moments accounting as proposed by Abadi et al. (2016) and calculates the privacy loss.

3.3 Evaluating Synthetic Data

Once the data has been generated, it is of course key to understand how close the synthetic data matches the original data. There are two ways to think about the general utility of this synthetic data. General utility compares the distribution between original data and synthetic data. Specific utility considers whether the results of particular analysis are similar.

General Utility. Snoke et al. (2018) suggest a measure for the general utility of synthetic data. They use the propensity score mean-squared error (pMSE) to measure differences between original and synthetic data. Original data and synthetic data are matched using propensity scores with the membership in the respective data sets as outcome variables. A classification model then predicts the provenance. The pMSE is therefore a measure of how good a classifier can distinguish real from synthetic observations. In principle, any prediction model could be used to discriminate between real and synthetic data. We employ the tree-based classification and regression trees (CART) from the synthpop R-package [15].

Theoretically a pMSE of 0 indicates that the synthetic data is the real data. While this is of course not desirable for synthetic data that intends to preserve privacy, the pMSE should still be as low as possible so that statistical analyses on the synthetic data are unbiased and efficient. Snoke et al. (2018) therefore also show how calculate the Null distribution, introducing the pMSE ratio and the standardized pMSE score as two additional utility measures. Similar to a t-statistic, the standardized value calculates the deviation from the null value. In practice however the ratio pMSE score seems to be more appropriate since the standardized value is oversensitive to small differences. We will report all three values in our assessment of the general utility of the synthetic data.

Specific Utility. While a notion of general utility is useful to assess synthetic data, social scientists are usually interested in the utility of specific queries to the data like, for example, regression coefficients. The specific utility of synthetic data is high when the regression estimates on the original data and the synthetic data are similar both in terms of bias and efficiency. In our experiments, we also assess the specific utility of the synthetic data. First, we compare correlation structures of the real data and synthetic data. Second, we also attempt to recover the regression coefficients of the regression in Equation 2.

[11] ReLUs (Rectified Linear Unit) are a non-linear function of x such that: $f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$.

[12] The sigmoid function – or also logistic function – transforms the input x such that: $f(x) = \frac{1}{1+e^{-x}}$.

[13] In theory gradients live on the real line. Without gradient clipping, the variance of the noise added to ensure DP would be infinitely large.

[14] More on using tensorflow for differential privacy can be found here: `https://medium.com/tensorflow/introducing-tensorflow-privacy-learning-with-differential-privacy-for-training-data-b143c5e801b6`

[15] Note that since the calculation of utility measures needs access to the real data the calculation of these measures is not differentially private.

Table 1: General Utility Measures for Synthetic Data with 5 Features.

	pMSE	Standardized Score	Ratio Score	pMSE	Standardized Score	Ratio Score	pMSE	Standardized Score	Ratio Score
	no DP	no DP	no DP	$\epsilon = 4$	$\epsilon = 4$	$\epsilon = 4$	$\epsilon = 1$	$\epsilon = 1$	$\epsilon = 1$
m = 1	0.10	0.73	1.04	0.14	8.37	1.51	0.18	21.50	1.93
m = 2	0.10	0.25	1.02	0.15	9.86	1.57	0.18	13.97	1.91
m = 3	0.10	1.31	1.08	0.14	10.44	1.50	0.17	9.82	1.89
m = 4	0.10	0.33	1.02	0.14	8.33	1.53	0.18	15.64	1.87
m = 5	0.09	-0.63	0.96	0.15	7.52	1.56	0.19	13.25	1.98
m = 6	0.09	-0.10	0.99	0.15	10.78	1.57	0.18	14.93	1.98
m = 7	0.10	0.58	1.03	0.14	9.53	1.44	0.18	14.00	1.89
m = 8	0.09	-0.60	0.96	0.14	6.68	1.54	0.18	14.46	1.97
m = 9	0.10	0.01	1.00	0.14	10.69	1.51	0.17	12.94	1.81
m = 10	0.09	-0.28	0.99	0.14	9.59	1.51	0.18	13.08	1.91
Average	0.10	0.16	1.01	0.14	9.18	1.53	0.18	14.36	1.91

4 Experiments

For the experimental setup of the simulation study, we look at several conditions as a combination of varying levels of privacy protection and varying number of features (variables). Overall, we define 6 different experimental conditions: three different levels of privacy × two different number of features.

We use the following levels of privacy protection in the experiment: no-DP, $\epsilon = 4$ and $\epsilon = 1$. Recall that lower levels of ϵ provide stronger privacy guarantees, but also need more noise during the training of our DP-GAN. Using a no-DP GAN and DP-GANs with varying ϵ allows us to evaluate the trade-off between data utility and privacy protection. For DP-synthetic data the choice of ϵ is not trivial and a "social question". The data owner has to decide on how much privacy loss is acceptable. Yet, a data release is only useful if some of the statistical utility is preserved. Complicating matters is that there is no straightforward interpretation of ϵ in the context of synthetic data. We chose the ϵ values in line with conventions in the DP literature. An information release worth $\epsilon = 1$ is considered as being an acceptable risk. In contrast, $\epsilon = 4$ would entail already a considerable privacy risk as it implies that the addition or removal of any observation in the original data changes the probability that the randomized synthesizer produces any possible outcome by a factor of $e^4 = 54.6$.

Our experimental setup also uses two different number of variables. the full data set consisting of all 41 features and a reduced data set only consisting of the five variables finally used in the regression. The varying number of features is implemented to evaluate how the dimensionality of the original data set affects the quality of the produced synthetic samples.

4.1 General Utility

We first evaluate the general utility of the data.

Synthetic Data with 5 features. Table 1 summarises our finding for a synthetic dataset with five features only. To represent uncertainty, we generate 10 synthetic data sets each. Every row of table 1 reports measures for the general utility of a synthetic data set for the experiment without differential privacy, at a fairly lose and privacy revealing setting for differential privacy ($\epsilon = 4$) and finally at an acceptable level of privacy protection with $\epsilon = 1$. The last line reports averages across all 10 data sets.

We find that the pMSE score for synthetic data without DP is fairly low and stable – indicating high general utility. As expected, the standardised score reports quite some variation. But again on average, the score is quite low. The ratio score without DP also indicates a very high utility of synthetic data. For $\epsilon = 4$, the pMSE score drops slightly to 0.14 on average. The standardised score increases quite substantially as does its variance. The ratio score increases up to 1.53 on average. Finally, at the highest privacy preserving value of $\epsilon = 1$, the pMSE score for synthetic data is at 0.18 on average. The standardised score goes further up, as does the ratio score.

Overall, our experiment shows that without privacy protection, the general utility of the synthetic data generated with the GAN is comparable with other data generating algorithms. In line with expectations, at higher levels of privacy protection, the general utility of the data goes down.

Synthetic Data with 41 features. Table 2 reports general utility measures for synthetic data with 41 variables. The layout of the table remains the same. Without differential privacy, the pMSE score now averages 0.18. The standardised score is at 11.05 and the ratio score at 1.37. At the first ϵ-level of $\epsilon = 4$, the pMSE score is at 0.21, the standardised score at 16.29 and the ratio score at 1.64. Finally, at an acceptable $\epsilon = 1$, the pMSE score is 0.23 on average, at 21.71

Table 2: General Utility Measures for Synthetic Data with 41 Variables.

	pMSE no DP	Standardized Score no DP	Ratio Score no DP	pMSE $\epsilon = 4$	Standardized Score $\epsilon = 4$	Ratio Score $\epsilon = 4$	pMSE $\epsilon = 1$	Standardized Score $\epsilon = 1$	Ratio Score $\epsilon = 1$
m = 1	0.18	10.83	1.33	0.22	20.88	1.67	0.23	24.21	1.82
m = 2	0.19	11.73	1.41	0.21	17.30	1.63	0.23	21.08	1.78
m = 3	0.18	12.23	1.37	0.21	15.68	1.64	0.23	19.49	1.78
m = 4	0.19	14.87	1.42	0.21	14.26	1.61	0.22	19.52	1.73
m = 5	0.19	10.98	1.38	0.21	17.53	1.66	0.23	18.87	1.76
m = 6	0.18	9.78	1.31	0.21	17.06	1.66	0.23	23.58	1.79
m = 7	0.19	10.88	1.37	0.21	17.61	1.63	0.24	20.33	1.85
m = 8	0.18	8.35	1.35	0.21	16.19	1.67	0.24	25.62	1.83
m = 9	0.19	10.37	1.38	0.21	11.96	1.65	0.23	22.49	1.77
m = 10	0.19	10.49	1.38	0.21	14.44	1.63	0.23	21.90	1.78
Average	0.18	11.05	1.37	0.21	16.29	1.64	0.23	21.71	1.79

for the standardised score and 1.79 for the ratio score. Again, as expected the general utility measures decline with increasing levels of data protection.

When comparing the general utility measures for synthetic data with five variables for those with 41 variables, the increase of data protection has different effects on the general utility. In principle the GAN has a harder time to synthesise data on the basis of a data set with 41 features when compared to only five features.[16] If there is no differential privacy, the pMSE for 5 variables is at 0.10, in contrast to a pMSE score without DP at 0.18 for 41 variables. However, adding differential privacy makes more of a difference to the general utility of data in the case of data with five variables, only. The general utility declines by 80% for synthetic data with five variables. However, it only declines by 27.78% for the data with 41 features, yet starting at a significantly worse no DP baseline.

4.2 Specific Utility

We now turn to evaluating the specific utility of synthetic data. We first consider the correlation structure within the data and compare the real data with synthetic data without DP, synthetic data at Epsilon equals four and epsilon equals one. We finally also compare regression coefficients of real and can synthetic data sets with five features and 41 features at different levels of privacy protection.

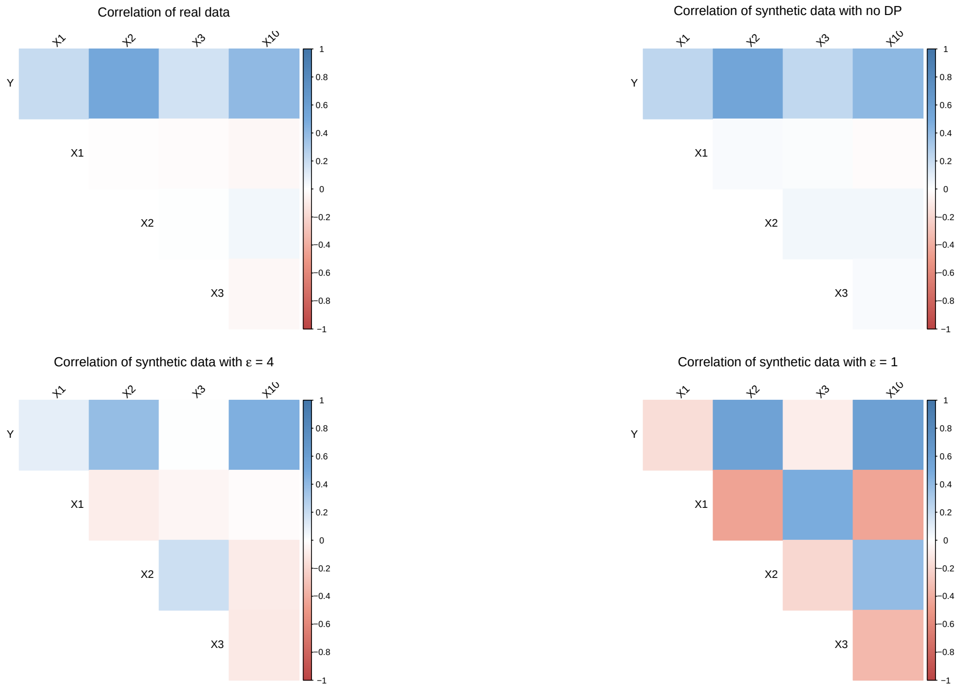
Synthetic Data with 5 features. Figure 1 shows the correlation structure among the data sets. The original data is in the upper left corner. We observe that the outcome variable Y in the real data correlates positively with all four explanatory variables. It has a medium correlation with X1 and X3 and correlates quite strongly with X2 and X10. There is almost no correlation amongst the three explanatory variables themselves. The upper right corner displays data on synthetic data with no DP. Here, the algorithm correctly replicates the correlation structure in the original data. The correlation among the three explanatory variable seems to be slightly stronger, but at negligible levels. The lower left corner shows the correlation of the synthetic data set at $\epsilon = 4$ which reveals quite a bit of information. The correlation between the outcome variable and X1, X2 and X10 is reproduced fairly well. Only the variable X3 now does not correlate any more with the outcome. The correlation among the explanatory variables themselves is already testament to the privacy protection imposed by the noise. Finally, the synthetic data in the lower right corner displays the correlation structure of synthetic data at an acceptable DP level of $\epsilon = 1$. While the correlation between X2 and X 10 is still retained with the outcome variable, all other correlations are at very different levels.

Figure 2 shows the regression coefficients from 10 synthetic data sets with different levels of privacy loss. The red square indicates the "true" regression coefficients calculated on the original data. We observe that the GAN captures the regression coefficients fairly well when using synthetic data without differential privacy. Only the intercepts tends to be distinct. At the $\epsilon = 4$, the algorithm is fairly off when estimating the regression coefficients for X3 and X 10. Surprisingly, in this experiment at $\epsilon = 1$ the regressing coefficients are pretty spot on — it seems to be even slightly better than for the synthetic data without differential privacy[17].

Synthetic Data with 41 features. For the synthetic data with 41 features we again begin with a closer look at the bi-variate correlations. The upper left corner with the original data shows that there are very few strong correlations in the original data. Only X1, X2, X3 and X10 positively correlate with the outcome variable. There is a substantive negative correlation between variables X6, X7 and X8. The upper right corner shows synthetic data without DP. Again, the algorithm manages to successfully reproduce most of the real correlations from the original data. The overall picture

[16] Note that until now we use a very simple baseline architecture for the GAN with the exact same architecture for the different numbers of features. We intend to engage in further testing of GAN architectures.

[17] This might be due to sheer luck. We will repeat the experiments many times to become more confident about our results.



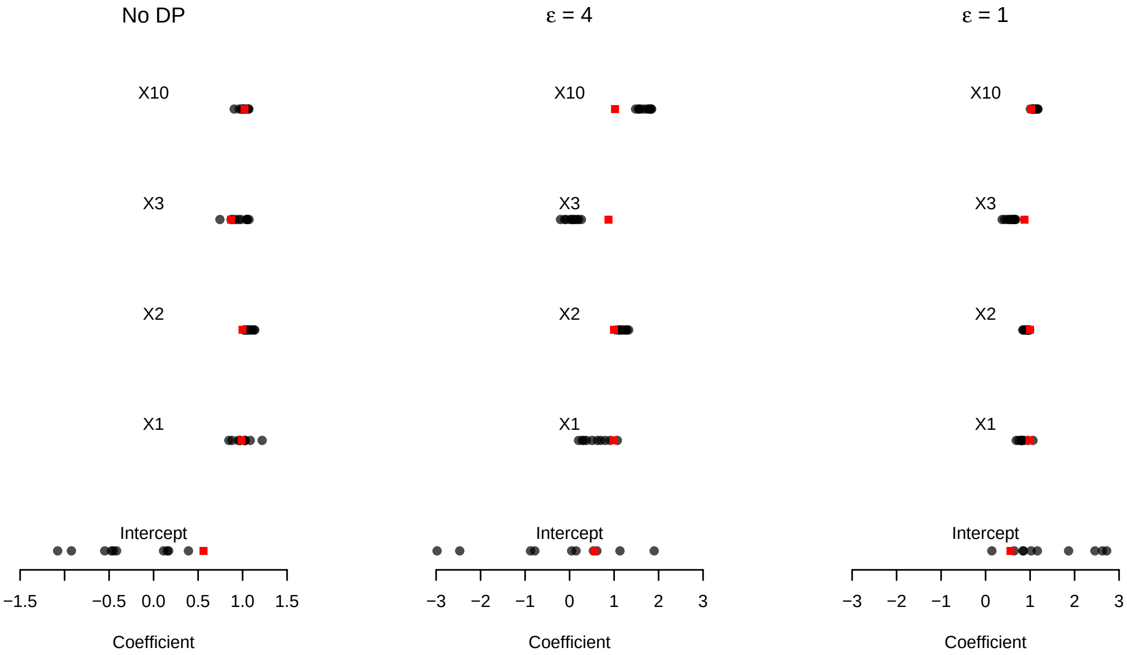
Correlation of real data

Correlation of synthetic data with no DP

Correlation of synthetic data with ε = 4

Correlation of synthetic data with ε = 1

Figure 1: Correlation Structure of Real and Synthetic Data with 5 features.



No DP

ε = 4

ε = 1

Figure 2: Regression Coefficients of real and 10 synthetic data sets with 5 features.

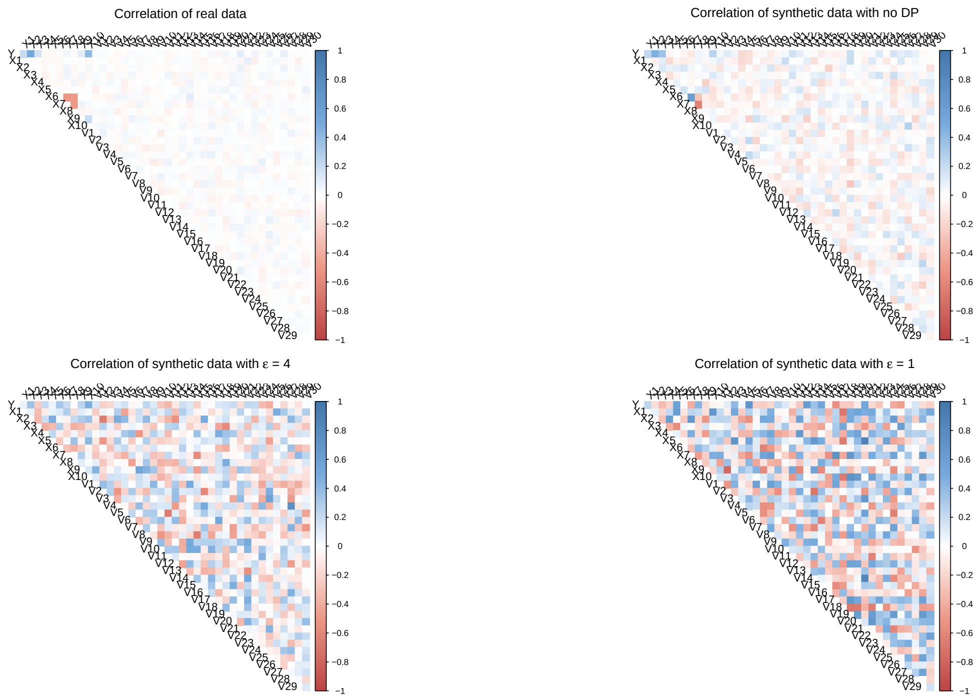


Figure 3: Correlation structure of real and synthetic data with 41 features.

— as was already an impression from the general utility — shows that the algorithm creates quite some noise inducing correlations for the synthetic data where there are none in the original data. The synthetic data at $\epsilon = 4$ clearly shows more noise. The plot for synthetic data at $\epsilon = 1$ is even more colourful — a clear testament to even more noise that has been added.

Figure 4 plots the regression coefficients from 10 synthetic data sets with different levels of privacy loss. As before, the red square represents the regression "true" coefficients calculated on the original data, the black dots stand for the regression coefficients estimated for each of the 10 synthetic data sets. Please note that the difference to figure 2 is not in the regression model – again we only calculate the regression equation 2. But we used the whole original data set to create our synthetic copies. The synthetic data without differential privacy also has high specific utility for this regression – the coefficients are fairly close to the "true" coefficients. Adding differential privacy to the data set with 41 features, however, severely harms the specific utility of the synthetic data for the regression task[18].

5 Conclusion and Outlook

In this paper we systematically evaluate DP-GANs from a social science perspective. We show that GANs produce synthetic data of high quality. The question we are interested in is how this utility is affected when generating synthetic data under differential privacy using a fairly simple DP-GAN.

The goal of this paper is to spark a conversation and exchange between Social Scientists and Computer Scientists about useful privacy preserving synthetic micro data. We show Social Scientists how to generate synthetic micro-data with privacy guarantees. We also turn to Computer Scientists. In highlighting the limitations of generating differentially private social science data, we intend to point to avenues for future research.

In particular we want to draw the attention to the following key findings. We observe that the utility – both general and specific – of synthetic data generated with a DP-GAN is significantly worse when the dimensionality of the data increases. We think that this issue could be potentially mitigated by carefully choosing the GAN architecture specific to the data that is to be synthesized. This means that Computer Scientists need to work closely with Social Scientists with domain knowledge of the data to construct better and problem-specific GAN architectures.

[18] Note that the scale of the x-axis in the three panels is quite different.

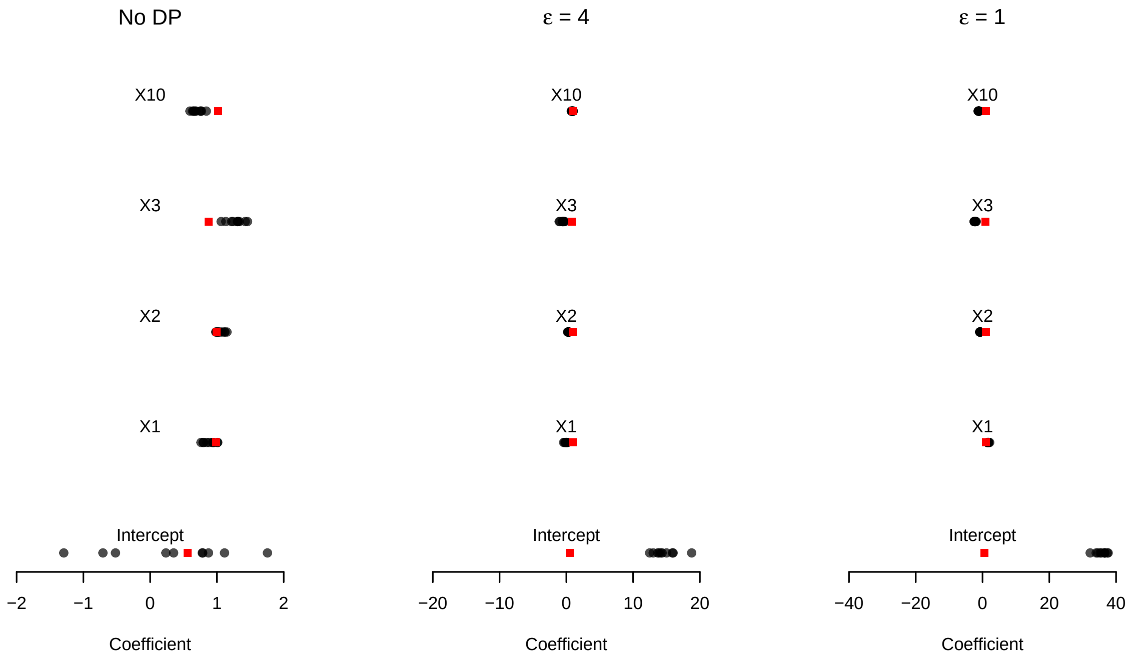


Figure 4: Regression Coefficients of Real and 10 Synthetic Data Sets with 41 Features.

In general, more research on the boundaries of utility and privacy is needed. It is clear that utility from a Social Scientist's perspective often implies specific utility for some class of models or hypotheses they intend to test with the data at hand. Again close collaboration of Computer Scientists and Social Scientists is needed to identify what defines the specific utility of a particular data set. This might be very different from data set to data set.

Further research on both sides also needs to consider other social science problems that were beyond the scope of this first experimental evaluation. Some of the open questions include: How to deal with missing data prior to synthesizing data? What about structural zeros in the data (e.g. if gender male, then status pregnant is impossible)? We hope that this paper can serve as the starting point for a fruitful debate and development of better privacy preserving synthetic data for social scientists.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep Learning with Differential Privacy. In *Proceedings of the 23rd ACM Conference on Computer and Communications Security*, Vienna, Austria, 2016. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL `https://arxiv.org/pdf/1607.00133.pdfhttp://arxiv.org/abs/1607.00133{%}0Ahttp://dx.doi.org/10.1145/2976749.2978318`.

Abowd, J. M. and Lane, J. New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers. In Domingo-Ferrer, J. and Torra, V. (eds.), *Privacy in Statistical Databases*, volume 3050, pp. 282–289. 2004. ISBN 978-3-540-22118-0. doi: 10.1007/978-3-540-25955-8_22.

Abowd, J. M. and Schmutte, I. M. An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1):171–202, 2019.

Abowd, J. M. and Woodcock, S. D. Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data. In *Privacy in Statistical Databases*, volume 3050, pp. 290–297. 2004. ISBN 3-540-22118-2. doi: 10.1007/978-3-540-25955-8. URL `http://www.scopus.com/inward/record.url?eid=2-s2.0-35048847594{&}partnerID=40{&}md5=bf5d85c779004da93401a0914af7acf2`.

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., and Greene, C. S. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. *bioRxiv*, pp. 159756, 2017. ISSN 15734951. doi: 10.1101/159756. URL `http://dx.doi.org/10.1101/159756http://www.biorxiv.org/content/biorxiv/early/2017/07/05/159756.1.full.pdf{%}0Ahttps://www.biorxiv.org/content/early/2017/07/05/159756{%}0Ahttps://www.biorxiv.org/content/early/2017/07/05/159756.full.pdf+html`.

Bellovin, S. M., Dutta, P. K., and Reitinger, N. Privacy and Synthetic Datasets. *SSRN Electronic Journal*, 1:1–52, 2018. doi: 10.2139/ssrn.3255766.

Caiola, G. and Reiter, J. P. Random Forests for Generating Partially Synthetic, Categorical Data. *Transactions on Data Privacy*, 3(1):27–42, 2010. ISSN 18885063. URL `http://www.tdp.cat/issues/tdp.a033a09.pdf`.

Drechsler, J. and Reiter, J. P. Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105(492):1347–1357, 2010. ISSN 01621459. doi: 10.1198/jasa.2010.ap09480. URL `http://www.tandfonline.com/action/journalInformation?journalCode=uasa20`.

Drechsler, J. and Reiter, J. P. An Empirical Evaluation of Easily Implemented, Nonparametric Methods for Generating Synthetic Datasets. *Computational Statistics and Data Analysis*, 55(12):3232–3243, dec 2011. ISSN 01679473. doi: 10.1016/j.csda.2011.06.006. URL `www.elsevier.com/locate/csdahttps://www.sciencedirect.com/science/article/pii/S0167947311002076?via{%}3Dihub`.

Dwork, C. Differential Privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052, pp. 1–12, Venice, Italy, 2006. Springer Verlag. ISBN 3-540-35907-9. URL `https://www.microsoft.com/en-us/research/publication/differential-privacy/https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/dwork.pdf`.

Dwork, C. and Roth, A. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2013a. ISSN 1551-305X. doi: 10.1561/0400000042. URL `https://www.cis.upenn.edu/{~}aaroth/Papers/privacybook.pdfhttp://www.nowpublishers.com/articles/foundations-and-trends-in-theoretical-computer-science/TCS-042`.

Dwork, C. and Roth, A. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–407, 2013b. ISSN 1551-305X. doi: 10.1561/0400000042. URL `https://www.cis.upenn.edu/{~}aaroth/Papers/privacybook.pdf`.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the Third Theory of Cryptography Conference*, pp. 265–284, 2006. ISBN 3-540-32731-2. doi: 10.1007/11681878_14. URL `http://link.springer.com/10.1007/11681878{_}14`.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, 2014. ISSN 10495258. doi: 10.1017/CBO9781139058452. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

Kinney, S. K., Reiter, J. P., and Berger, J. O. Model Selection When Multiple Imputation is Used to Protect Confidentiality in Public Use Data. *Journal of Privacy and Confidentiality*, 2(2):3–19, 2010. URL http://repository.cmu.edu/cgi/viewcontent.cgi?article=1002{&}context=jpc{_}forthcoming.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. *International Statistical Review*, 79(3): 362–384, 2011. ISSN 03067734. doi: 10.1111/j.1751-5823.2011.00153.x.

Little, R. J. Statistical Analysis of Masked Data. *Journal of Official Statistics*, 9(2):407–426, 1993.

Montgomery, J. M. and Olivella, S. Tree-Based Models for Political Science Data. *American Journal of Political Science*, 62(3):729–744, 2018. ISSN 00925853. doi: 10.1111/ajps.12361. URL http://dx.doi.org/10.7910/.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *Proceedings - IEEE Symposium on Security and Privacy*, pp. 111–125, 2008. ISBN 9780769531687. doi: 10.1109/SP.2008.33. URL https://www.cs.utexas.edu/{~}shmat/shmat{_}oak08netflix.pdf.

Nissim, K., Steinke, T., Wood, A., Bun, M., Gaboardi, M., O 'brien, D. R., and Vadhan, S. Differential Privacy: A Primer for a Non-technical Audience * (Preliminary version). (1237235), 2017. URL http://privacytools.seas.harvard.eduhttp://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp{_}0.pdf.

Page, H., Cabot, C., and Nissim, K. Differential privacy: an introduction for statistical agencies. *National Statistician's Quality Review into Privacy and Data Confidentiality Methods*, (December):1–53, 2018.

Raghunathan, T. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16, 2003. doi: 10.1007/s13398-014-0173-7.2. URL http://www2.stat.duke.edu/courses/Spring06/sta395/raghunathan2003.pdfhttp://hbanaszak.mjr.uw.edu.pl/TempTxt/RaghunathanEtAl{_}2003{_}MultipleImputationforStatisticalDisclosureLimitation.pdf.

Reiter, J. P. Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21(3): 441–462, 2005. URL http://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/using-cart-to-generate-partially-synthetic-public-use-microdata.pdffiles/2405/Reiter-2003-UsingCARTtogeneratepartiallysynthetic,public.pdf{%}5Cnfiles/2397/summary.html.

Reiter, J. P. and Raghunathan, T. E. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471, 2007. ISSN 01621459. doi: 10.1198/016214507000000932. URL https://www.tandfonline.com/doi/pdf/10.1198/016214507000000932?needAccess=true.

Rubin, D. B. Discussion: Statistical Disclosure Limitation, 1993.

Snoke, J., Raab, G., and Slavkovic, A. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 2018. ISSN 09641998. doi: 10.1111/rssa.12358. URL http://arxiv.org/abs/1604.06651.

Sweeney, L. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997.

Triastcyn, A. and Boi Faltings. Generating Artificial Data for Private Deep Learning. 2018. URL https://arxiv.org/pdf/1803.03148.pdf.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. Differentially Private Generative Adversarial Network. 2018. URL https://doi.org/10.475/123{_}4http://arxiv.org/abs/1802.06739.

Xu, C., Ren, J., Zhang, D., Zhang, Y., Qin, Z., and Ren, K. GANobfuscator: Mitigating Information Leakage under GAN via Differential Privacy. *IEEE Transactions on Information Forensics and Security*, 14(9):1–1, 2019. ISSN 1556-6013. doi: 10.1109/tifs.2019.2897874.