Epistemic Parity: Reproducibility as an Evaluation Metric for Differential Privacy

Lucas Rosenblatt[*]
New York University
New York, NY, USA
lucas.rosenblatt@nyu.edu

Bernease Herman
University of Washington
Seattle, WA, USA
bernease@uw.edu

Anastasia Holovenko
Ukrainian Catholic University
Lviv, Ukraine
anastasia.holovenko@ucu.edu.ua

Wonkwon Lee
New York University
New York, NY, USA
wl2733@nyu.edu

Joshua Loftus
London School of Economics
London, UK
J.R.Loftus@lse.ac.uk

Elizabeth McKinnie
Microsoft
Seattle, WA, USA
Elizabeth.McKinnie@microsoft.com

Taras Rumezhak
Ukrainian Catholic University
Lviv, Ukraine
rumezhak@ucu.edu.ua

Andrii Stadnik
Ukrainian Catholic University
Lviv, Ukraine
andrii.stadnik@ucu.edu.ua

Bill Howe
University of Washington
Seattle, WA, USA
billhowe@uw.edu

Julia Stoyanovich
New York University
New York, NY, USA
stoyanovich@nyu.edu

ABSTRACT

Differential privacy (DP) data synthesizers are increasingly proposed to afford public release of sensitive information, offering theoretical guarantees for privacy (and, in some cases, utility), but limited empirical evidence of utility in practical settings. Utility is typically measured as the error on representative proxy tasks, such as descriptive statistics, multivariate correlations, the accuracy of trained classifiers, or performance over a query workload. The ability for these results to generalize to practitioners' experience has been questioned in a number of settings, including the U.S. Census. In this paper, we propose an evaluation methodology for synthetic data that avoids assumptions about the representativeness of proxy tasks, instead measuring the likelihood that published conclusions would change had the authors used synthetic data, a condition we call epistemic parity. Our methodology consists of reproducing empirical conclusions of peer-reviewed papers on real, publicly available data, then re-running these experiments a second time on DP synthetic data and comparing the results.

We instantiate our methodology over a benchmark of recent peer-reviewed papers in the social sciences. We express the authors' claims computationally to automate the experiment, generate DP synthetic datasets using multiple state-of-the-art mechanisms, then estimate the likelihood that these conclusions will hold. We find that, for reasonable privacy regimes, DP synthesizers can achieve high epistemic parity for several papers in our benchmark. However, some papers, and particularly some specific findings, are difficult to reproduce for any of the synthesizers. Given these results, we recommend a new class of mechanisms that offer stronger utility guarantees (as measured by epistemic parity) and more nuanced privacy protection using application-specific risks and threat models.

1. INTRODUCTION

Differential privacy (DP) has been studied intensely for over a decade, and has recently enjoyed uptake in both the private and public sectors. In situations where the downstream analysis is known, one can design specialized mechanisms with high utility [37, 38]. But an active research area is to design general DP data synthesizers (henceforth, synthesizers) that model the entire data distribution, inject noise, then sample the noisy model to generate synthetic datasets intended to be broadly usable in a variety of unanticipated applications. Evidence to support claims of general utility is typically presented as results on proxy tasks over common public datasets (e.g., the ubiquitous Adult dataset [33]). Proxy tasks may include descriptive statistics, queries involving one or two variables [25, 24, 50, 51], classification accuracy [12, 50, 56], and information theoretic measures [56]. Although these proxy tasks are procedurally representative of real tasks, the implicit claim of generalization to practice is rarely explored.

Limited empirical evidence on relevant tasks undermines trust in the practical use of DP. The US Census Bureau adopted DP for disclosure avoidance in the 2020 census, interpreting federal law (the Census Act, 13 U.S.C. § 214, and the Confidential Information Protection and Statistical Efficiency Act of 2002) as a mandate to use advanced methods

[*] Rosenblatt is the first author, Howe and Stoyanovich are the senior authors.

Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment. This is a minor revision of the paper entitled "Epistemic Parity: Reproducibility as an Evaluation Metric for Differential Privacy," published in PVLDB, Vol. 16, No. 11, ISSN 2150-8097. DOI: https://doi.org/10.14778/3611479.3611517
.

to protect against computational reconstruction attacks unforeseen when the laws were passed. But the adoption of DP for the Census was met with resistance among many in the research community, who contend that data infused with DP noise affects demographic totals [47] and exacerbates underrepresentation of minorities [32, 21]. Besides the research implications, there are potential consequences for policy: Block grants are allocated based on minority populations as measured by the census data, and underrepresentation can lead to underfunding integral services including Medicaid, Head Start, SNAP, Section 8 Housing vouchers, Pell Grants, and more [8]. Although the Census Bureau held workshops, released demonstration datasets, and published technical reports to support the community, these outreach efforts realized limited success; multiple lawsuits are still pending as of May 2023.

Despite these challenges, DP still offers stronger guarantees of disclosure protection than, and similar utility to, alternative proposals (e.g., k-anonymity, swapping [8]). DP, when used correctly, ensures that any inferences conducted on data do not reveal whether a single individual's information (including, for example, their gender or race) was included in the data for analysis [15]. DP can therefore not only protect privacy, but also enable access to protected demographic attributes necessary for research on fairness and equity in machine learning [29].

Characterizing DP Error.

A practical method of operationalizing DP is to learn a (noise-infused) model of a dataset, then sample that privatized model to generate synthetic data that can be released publicly [16, 22, 52, 45, 54, 37]. Ideally, this approach would provide a drop-in replacement for the original data that can be used in *any* downstream context to produce reasonably faithful results with strong privacy guarantees. But this ideal is unrealizable, both theoretically and practically. Overly accurate estimates of too many statistics are blatantly non-private, affording full reconstruction of the original dataset [13]. For any DP synthetic dataset, some statistics will tend to be faithful to the original data, while others will incur essentially arbitrary error. If the privacy budget is allocated uniformly across features, descriptive statistics of each individual feature will be faithful, but the conditional probabilities and marginals needed to construct the joint probability distribution, which is needed for general inference, will be unreliably noisy, and vice versa. Utility loss may also be non-uniform across subsets of a dataset, in some cases exacerbating inequity and leading to underrepresentation [2, 32] or to error rate disparities [43]. Designers of DP synthesizers must therefore make some kind of educated guess about which tasks should be preserved and which can be ignored. Further, the error introduced by DP methods can and should be incorporated into statistical models explicitly, just as other sources of error are modeled explicitly. However, current DP synthesizers tend not to provide formal descriptions of the error they introduce; this lack of error guarantees is a major drawback of private data release. Our work does not address this limitation, but does help provide an empirical motivation for doing so.

Methodology.

We propose an evaluation methodology for DP synthesizers based on reproducibility: that *published findings on the original dataset should be replicable on a noise-infused dataset.* We identify conclusions in the text of published papers, extract relevant findings supporting those conclusions, implement the corresponding statistical tests using the authors' data, generate synthetic datasets using state-of-the art DP synthesizers, re-apply the statistical tests over the synthetic data, and then determine if the findings still hold. If all findings hold, we say that the DP synthesizer achieves *epistemic parity* for that paper.

We instantiate our methodology over a benchmark of peer-reviewed sociology papers that are based on public data from the Inter-university Consortium for Political and Social Research (ICPSR) repository. We model quantitative results as an inequality between two numbers, for example, "Those using marijuana first (vs. alcohol or cigarettes first) were more likely to be Black, American Indian/Alaskan Native, multiracial, or Hispanic than White or Asian." [19]

Following Errington *et al.* [18], and as is common in the reproducibility literature, our aim was to identify and reproduce a selection of key findings from each paper. For generality, interpretability and simplicity, we consider whether a conclusion holds over synthetic data to be true if the the two quantities are in the same relative order, and do not attempt to measure the change in effect size or the statistical significance of the difference between the original and synthetic result.

Benchmark and results.

ICPSR is an NSF-funded repository for social science data holding over 100,000 publications associated with 17,312 studies. A study typically involves hundreds of variables and supports dozens of papers. Each paper can be considered to be deriving its own dataset (selected variables and selected rows) from the source data of the study. We apply DP methods to synthesize data for these paper-specific, study-derived datasets. ICPSR studies are publicly available by policy, which enables us to instantiate the epistemic parity methodology and develop a benchmark. Notably, there is increasing demand from the ICPSR leadership and community to support keeping sensitive data private, while generating DP synthetic subsets to support reproducibility. Our methodology can be used to respond to this demand.

Paper selection. The benchmark consists of 4 datasets and 8 recent peer-reviewed papers selected for impact, accessibility of the topic to non-experts, recency, and several other criteria. We extracted findings and attempted to reproduce them, following the "same data, different code, different team" approach to reproducibility, encountering challenges commonly reported in that literature including undocumented data versioning, unspecific or incomplete methodologies, and irreconcilable differences between our reproduction and what the authors report. A complete list of papers that we attempted to reproduce, and the issues we encountered, is available in our public GitHub repository. [1]

DP synthesizer selection. We use six state-of-the-art DP synthesizers, namely, MST [37], PrivBayes [56], PATECT-GAN [45], AIM [38], PrivMRF [7], and GEM [35] executing each at their recommended settings.

Summary of results. We find that marginals-based and Bayes-net based state-of-the-art DP synthesizers are able to achieve high epistemic parity for five out of eight papers in

[1] https://github.com/DataResponsibly/SynRD

our benchmark, but that some papers, and particularly some specific findings, are difficult to reproduce for any of the synthesizers, suggesting a basis for a new benchmark. The papers on which high epistemic parity is achieved use relatively low-dimensional tabular data. However, as we show empirically, large domain and high-dimensional settings are still a bottleneck for increased adoption of DP synthesizers.

Roadmap and Contributions.

We discuss background and relevant DP synthesis methods in Section 2, and then present our contributions: *(i)* the epistemic parity evaluation methodology, based on reproducing qualitative and quantitative empirical findings in peer-reviewed papers over DP synthetic datasets (Section 3); *(ii)* an instantiation of the methodology for eight peer-reviewed social science publications, creating a reusable benchmark for evaluating synthesizers (Section 4); and *(iii)* an experimental evaluation on our benchmark, using five state-of-the-art DP synthesizers (Section 5). We conclude with a discussion of the results, identifying trade-offs and motivating a new class of privacy techniques that favor strong epistemic parity and de-emphasize privacy risk, in Section 6.

2. BACKGROUND

Differential privacy (DP) ensures that altering or removing one record from a given dataset does not significantly affect the outcome of an analysis or query. Intuitively, DP prevents an observer of a private output from drawing conclusions about which specific individuals' information was included in the input. DP is based on the concept of neighboring datasets, where two datasets are neighboring if they differ in a single record. In the scope of the private synthesizers considered by this paper, datasets X and X' are considered neighboring if the removal of a single element x_i from one yields the other (except in the case of PrivBayes; we account for this in our budget allocation). Informally, DP synthesis mechanisms ensure that a synthetic dataset derived from two neighboring datasets will be similar enough as to hide the presence or anbsence of the removed element.

Different mechanisms use different formulations of DP: AIM and GEM both give *concentrated differential privacy* (ρ-zCDP) guarantees [6], while MST, PATECTGAN and PrivMRF give conventional (ϵ, δ)-DP guarantees, and PrivBayes gives an $(\epsilon, 0)$-DP guarantee. As demonstrated by Bun*et al.* [6], an established hierarchy of these guarantees exists: an $(\epsilon, 0)$-DP mechanism gives $\frac{\epsilon^2}{2}$-zCDP, which gives $(\epsilon\sqrt{2\log(1/\delta)}, \delta)$-DP for every $\delta > 0$. In our experiments, all ϵ parameters are translated using these relationships so as to compare at the same relative privacy settings. As is typical, we set δ to be "cryptographically small:" at most $\frac{1}{n}$ for n records, but typically much smaller [17].

Differentially Private Data Synthesis.

We considered five state-of-the-art private data release methods: MST, AIM, PrivMRF, PATECTGAN, PrivBayes and GEM. We acknowledge that many other methods exist for generating DP data [16, 22, 52, 54]. We chose this set informed by recent work [51, 38] showing that, over randomized query workloads on tabular data, MST, AIM and PrivMRF are the highest-performing marginal-based methods, that PrivBayes is the highest-performing Bayes-net-based method, and that PATECTGAN and GEM are the highest-performing deep learning based methods. AIM, PrivMRF and GEM are more recent than MST; they were not included in the recent dedicated DP synthesizer benchmarking survey [51] and are currently considered to be the state-of-the-art DP synthesizers.

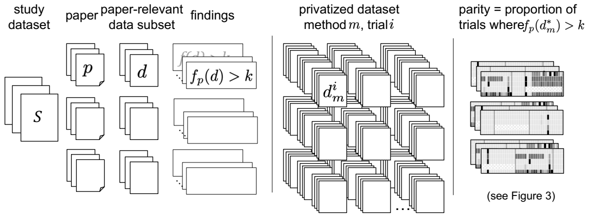
study dataset
paper
paper-relevant data subset
findings
privatized dataset method m, trial i
parity = proportion of trials where $f_p(d^*_m) > k$
S
p
d
$f(d) > k$
$f_p(d) > k$
d^i_m
...
(see Figure 3)

Figure 1: Epistemic parity workflow: Each study dataset supports many papers, each using a subset of the features. The paper's findings are implemented as computable inequalities. We generate many privatized datasets using different random seeds, then compute the proportion of these trials for which the findings hold (Figure 2).

PrivBayes [56] derives a Bayesian model and adds noise to all k-way correlations to ensure differential privacy, and despite being published in 2017 is still competitive with more recent methods. MST [36] relies on the Private-PGM graphical model to construct a maximum spanning tree among attributes in the data feature space, where edges are weighted by mutual information. It can measure 1-, 2- and even 3-way marginals to create a high-fidelity low-dimensional approximation of the joint distribution. AIM [38], like MST, relies on the Private-PGM for parameterizing the underlying distribution, but utilizes an iterative process to take advantage of higher values of ϵ. PrivMRF [7] is another marginal-based algorithm that relies on Private-PGM, and its novelty lies in a clever criteria for the selection of marginals to measure. PATECTGAN [55, 45] relies on a conditional generative adversarial network tuned to tabular data, where the discriminator has privacy constraints. GEM[35] analyzes the "Adaptive Measurements" framework for private synthetic data algorithms, inspired by the MWEM architecture [22], to (1) privately selects a set of queries; (2) obtains noisy measurements of these queries; and (3) updates an approximating distribution according to some loss function.

3. EPISTEMIC PARITY EVALUATION

Intuitively, epistemic parity holds if all published findings from the *original dataset* also hold on the *synthetic dataset*. Consider a finding to be a Boolean condition over the dataset, e.g., whether some statistic f exceeds threshold k. We obtain an epistemic parity score by synthesizing many datasets and reporting the fraction for which the finding holds. Figure 1 illustrates the workflow. The input is a set of papers, and the output is a set of scores indicating whether findings are supported under various DP synthesizers. A study is associated with one dataset and potentially many papers, each using a subset of the variables in the study. We assume public access to the data on which the paper's results were computed; our focus is on evaluating DP methods (requiring ground truth) rather than on protecting the privacy of subjects involved in the study.[2]

[2] Indeed, inaccessible ground truth undermined the US Cen-

Given a paper, we identify natural language claims made by the authors as candidates for findings. Though these claims may appear anywhere in the paper, most were found in the results section. Domain expertise provides an advantage in this task, but we contend that it should always be possible for non-expert readers to identify major claims since the goal of a paper is to communicate findings to a broader audience. For each claim, we identify the quantitative evidence that supports the claim, recording the variables involved and methods used. We then re-implement the analysis to (attempt to) reproduce the salient findings and conclusions in the paper over the original, public dataset.

While this reproduction step is always possible in principle, it can be difficult or impossible in practice [3, 39], and may involve guesswork when the computational details are incomplete. Moreover, inconsistent reproducibility can introduce bias in our benchmark: we may be more likely to include findings for which computational details are clear, which may be those that are simpler to explain or better-known by the author.

If the reproduction was successful, we generate $k \times m$ synthetic datasets representing k trials with different random seeds and m different DP methods, and then draw an additional B samples from each seeded DP method. In our initial benchmark, $k = 10$ and $m = 5$, and $B = 25$. The additional B draws allow us to bootstrap a confidence interval for each (trained) synthesizer. That is, there are two sources of randomness: the training procedure used by the mechanism, and the random sampling of the learned model to actually generate synthetic data. Although each synthetic dataset could be scaled to any number of records — recall that we are sampling a privatized model — we always use the same number of records as the original data for each bootstrap sample. Given this set of synthetic datasets, we again attempt to reproduce the findings using each one. Finally, we contrast the findings based on original and DP data by measuring the proportion of trials, for each method, where a given finding holds. Our methodology is implemented in an open-source framework.

Reproducing Experimental Studies.

We adapt three concepts of reproducibility—*values*, *findings*, and *conclusions*—from Cohen *et al.* [9] into a practical taxonomy for reproducing a statistical analysis in a peer-reviewed publication, and implement a software framework that allows us to conduct concrete experiments around this taxonomy. The atomic element in reproducibility is a *finding*, defined by Cohen *et al.* [9] as "a relationship between the *values* for some reported figure of merit with respect to two or more dependent variables." For the purposes of our study, a *finding* consists of a natural language statement (i.e., a *claim*) reported in a publication, along with evidence provided by one or more quantitative or qualitative sub-statements about the analysis.

Evidence for a *finding* consists of a comparison between two or more *values* that can be evaluated as a Boolean condition. A value may be a scalar (i.e., 34.1%), an aggregated or computed result (i.e., a regression coefficient of 1.2), or even an implicit threshold expressed in natural language (e.g., "a low rate" or "a strong correlation"). In these cases, we instantiate the language as a quantitative threshold, applying conventions from the literature when they exist. For example, a common convention is that Pearson's correlation is considered "strong" when r is larger than 0.7.

A special case of a *finding* is a qualitative *visual finding* that often appears in the form of a figure, table or diagram. A figure encodes many potential *findings*; we do not (necessarily) consider each of these sub-findings on their own in our analysis, but rather treat them as a single *visual finding*: we attempt to reproduce the figure itself, and subjectively evaluate its similarity to the original.

Finally, following Cohen *et al.* [9], a *conclusion* is defined as "a broad induction that is made based on the results of the reported research." A conclusion must be explicitly stated in a paper, and comprises one or several *findings*.

Generating DP Synthetic Data.

Each of the papers that we reproduced using DP synthetic data derived findings from a subset of the full study's data. For example, HSLS:09 consists of over 7000 columns, but Jeong *et al.* [30] used only a subset of 57. We synthesize the subset of data relevant for the reproduced findings and conclusions, as discussed in Section **??**. In the case where a paper relies on longitudinal data from a study, we collapse the data such that it is "one row to one person."

The DP methods for private data release are executed for the range of ϵ values $\epsilon \in \{e^{-3}, e^{-2}, e^{-1}, e^{0}, e^{1}, e^{2}\}$, which represents a small to medium privacy regime [4]. Each DP mechanism is run 10 times to produce, at each ϵ value, $10 \times B$ sampled datasets using the same sample size but different random seeds (where B is the bootstrap parameter). Each DP method involves different hyperparameters and varying levels of tunability, but we use author-recommended settings to avoid biasing results towards our own expertise. We then re-compute the findings for each sample.

If all findings are reproduced regardless of *epsilon* or random seed, we say that the DP mechanism achieves *complete* epistemic parity. But we measure parity as the *proportion* of iterations for which the finding holds. The goal is to overlook small variations in the exact value in favor of maintaining the relative relationships of the computed statistics for interpretability and practical utility.

4. BENCHMARK CONSTRUCTION

In constructing our benchmark, we selected study datasets that have been used in at least 100 papers, focusing on peer-reviewed, publicly available studies from the past 5 years that utilize publicly accessible data and are under 30 pages. We selected: (1) The High School Longitudinal Study (HSLS:09) [10], a longitudinal study of U.S. 9th graders (three of four paper reproduction attempts successful), (2) the National Longitudinal Study of Adolescent and Adult Health (AddHealth) [23] that follows U.S. adolescents from grades 7 through 12 during the 1994-1995 school year (two of four paper reproduction attempts successful), (3) The National Survey on Drug Use and Health (NSDUH) [53] that measures U.S. drug use prevalence and correlates (one of four paper reproduction attempts successful, at least partially due to study variations without clear version records), and (4) the Americans' Changing Lives Survey (ACL) [26], which tracks U.S. adults over time to understand the effects of social connections and work on health (two of two reproduction attempts partially successful), see [44] for details.

sus Bureau's efforts to build trust in DP [5].

Selected Studies.

A study dataset was selected only if it was used in at least 100 papers. For each selected study, we selected peer-reviewed papers published during the past 5 years that are no more than 30 pages long.

HSLS:09: High School Longitudinal Study [10], is a nationally representative, longitudinal study of U.S. 9th graders who were followed through their secondary and postsecondary years.

AddHealth, National Longitudinal Study of Adolescent and Adult Health [23], consists of a nationally representative sample of U.S. adolescents in grades 7 through 12 during the 1994-1995 school year.

NSDUH, National Survey on Drug Use and Health 2004-2014 [53], measures the prevalence and correlates of drug use in the U.S.

ACL, The Americans' Changing Lives Survey [26], is an ongoing longitudinal study of the lives of U.S. adults. The study has several waves, the first of which was conducted in 1986, and each wave continues with the same respondents to determine how social connections, work, and other factors affect health throughout their lifetimes.

Selected Papers.

We will briefly outline each paper from our benchmark. Saw *et al.* [48] utilized HSLS:09 for examining disparities in STEM career aspirations among high school students. Lee *et al.* [34] evaluated the impact of teacher support and self-perceptions on math performance using HSLS:09. Jeong *et al.* [30] investigated racial bias in the performance of machine learning classification tasks with HSLS:09. Fruiht and Chan [20] explored the impact of mentors on first-generation college students using AddHealth. Iverson and Terry [27] analyzed the effects of high school football on later-life depressive and suicidal tendencies using AddHealth. Fairman *et al.* [19] investigated early marijuana use and its consequences using NSDUH. Assari and Bazargan [1] studied the impact of obesity on mortality risk due to cerebrovascular disease using ACL. Pierce and Quiroz [40] examined the effects of social support on emotional states using ACL.

Note on study/dataset dimensionality. We did not explicitly filter papers based on the size of the dataset they used. The studies we considered were very high dimensional (many thousands of variables), but the corresponding papers in our benchmark each follow a standard subsetting procedure, where they select a small collection of variables of interest for analysis. Thus, our benchmark datasets are not as high-dimensional as other benchmarks [38].

Comparison to Other DP Benchmarks.

Selected papers represent 8 new datasets. In this section, we adopt a meta-learning perspective [41, 42] to discuss characteristics that differentiate these datasets from typical ML benchmarks [33, 49] used in prior DP studies [24, 45].

In Table 1, we show several properties and meta-features for eight datasets from our benchmark, as well as for two popular datasets from the UCI Machine Learning repository [14], Adult [33] and Mushroom [49].

Number of outliers is calculated as the number of values that fall outside of the second and third quartiles, summed across all numerical variables. Outliers present a challenge for privatization, as they are easily identifiable.

Mutual information (mean, standard deviation) is calculated for each pair of features. DP synthetic data algorithms like PrivMRF, MST, PrivBayes and AIM are, at their core, interested in *preserving* mutual information between features, but this preservation is challenging given the constrained nature of model fitting (often relying on a small set of 2- or 3-way marginal queries) and the addition of noise for privatization.

Skewness (mean, standard deviation) of a sample is calculated according to the formula for adjusted Fisher-Pearson standardized moment coefficient, which is an unbiased estimate that gives similar results to other popular skewness measures for large samples, but can vary for smaller and moderate-sized samples [31]. The regularity of variables in a dataset (the level of assymmetry in the underlying distributions) affects their ease of replication.

Sparsity (mean, standard deviation) is defined as a normalized ratio of the number of samples over the number of unique values. Sparser data may be harder to capture through noisy marginal measurements.

Table 1 illustrates the benchmark covers a wide range of values of these metrics. Interestingly, one of our most challenging datasets to reproduce, Iverson and Terry [27], had the lowest average mutual information score and one of the highest sparsity scores. Many of the synthesizers we test depend on mutual information to select the marginal measurements for distribution learning. Selecting the most relevant 2-way marginals when mutual information is uniformly *low* and there are many features is clearly a challenge. Moreover, Adult, a common challenge dataset, has uniquely skewed distributions, which aligns with prior work suggesting that this dataset is idiosyncratic therefore less appropriate for evaluation and benchmarking [11].

5. RESULTS

Our benchmark consists of eight papers, each evaluated on six synthetic data algorithms for six values of ϵ, for a total of 36 mechanisms for each paper, each repeated with 10 random seeds. We draw 25 samples of size n, where n is the real data sample size, and bootstrap over this set of samples when calculating average parity over our finding set. Benchmarking extensively with DP synthesizers is computationally expensive [38, 45, 51]. Fitting many synthesizers took 100s of compute hours. Training PrivMRF and PATE-CTGAN was done using NYU's Greene High Performance Computing cluster using A100 and RTX8000 NVIDIA GPUs with 80GB and 48GB of RAM respectively. CPUs from that same cluster were used to train AIM, MST, PrivBayes, and GEM. The benchmark itself (assessing parity per paper) was also run on the cluster.

Epistemic parity: overall performance.

Figure 2 shows parity for all findings across all papers, for each of the five synthesizers, with ϵ regimens of e^{-3}, e^{-2}, e^{-1}, e^{0}, e^{1}, and e^{2}. Darker color indicates lower average parity, while lighter indicates higher average parity. Each paper is a block of rectangles, where the x-axis represents *findings* and the y-axis shows the five synthesizers. The crosshatched cells indicate that a synthesizer was unable to fit to a dataset in under 6 hours.

The final row, labeled "real, bootstrap," in Figure 2 shows the results of our Bayesian bootstrapping control procedure (see full version of the paper for details [44]). We note that over 97% of our findings are reproduced in 100% of our

Table 1: Properties and meta-features of the datasets in our benchmark, and of two datasets that are commonly used for DP benchmarking, Adult and Mushroom. Mutual Information, Skewness and Sparsity are the *average* for each of these metrics across all variables in the dataset. Our results reinforce that synthesizers may struggle with large sample sizes (Fairman *et al.*), large domain sizes (Jeong *et al.*), and low mutual information (Iverson and Terry).

Paper	Sample Size	Variables	Domain Size	Outliers	Mutual Info.	Skewness	Sparsity
Assari and Bazargan [1]	3361	16	9.03e+09	9	0.051 ± 0.153	0.563 ± 1.557	0.253 ± 0.231
Fairman *et al.* [19]	**293581**	6	2.03e+05	0	0.255 ± 0.432	0.185 ± 0.462	0.174 ± 0.165
Fruiht and Chan [20]	4173	11	2.20e+05	6	0.104 ± 0.256	0.607 ± 1.694	0.394 ± 0.183
Iverson and Terry [27]	1762	27	5.71e+15	5	**0.004 ± 0.010**	NaN	0.307 ± 0.180
Jeong *et al.* [30]	15054	57	**7.04e+42**	32	0.020 ± 0.026	0.338 ± 2.850	0.261 ± 0.166
Lee *et al.* [34]	14575	9	5.11e+17	5	2.862 ± 1.242	0.080 ± 0.440	0.111 ± 0.156
Pierce and Quiroz [40]	1585	17	7.19e+11	11	0.030 ± 0.050	0.001 ± 1.050	0.146 ± 0.158
Saw *et al.* [48]	20242	9	4.30e+04	3	0.143 ± 0.145	1.291 ± 1.218	0.354 ± 0.171
Adult [33]	32561	15	9.06e+14	96	0.066 ± 0.053	17.455 ± 22.992	0.125 ± 0.164
Mushroom [49]	8124	23	2.44e+14	74	0.199 ± 0.209	6.211 ± 8.955	0.297 ± 0.219

Bayesian bootstrap iterations. For the remaining inconsistent three findings over the bootstrap, it is unfair to expect the private synthesizers to have higher epistemic parity than the bootstrap control.

The overall performance of the synthesizers was impressive. All synthesizers achieved 100% parity for Lee *et al.* [34], and Fruiht and Chan [20]. Besides PrivMRF (which was computationally infeasible to fit to the data), AIM, MST, PrivBayes, PATECTGAN, and GEM achieved 100% parity for Pierce and Quiroz [40] as well. Both Saw *et al.* [48], and Assari and Bazargan [1] also had very high levels of parity between findings on real and on synthetic data, although each of these papers had at least one finding that was difficult to reproduce.

Two of the papers provided the greatest challenge, and the most interesting results, across privacy regiments and synthesizer types: Fairman *et al.* [19], and Iverson and Terry [27]. These papers were challenging for very different reasons. Fairman *et al.* [19] had the second-smallest domain size, and the fewest variables. However, it had by far the largest sample, consisting of nearly 300K records. This combination made it very sensitive to noise in marginal measurements (as they are essentially counts), in turn making the findings difficult to replicate in low-privacy settings. Still, PrivBayes and MST exhibited impressive performance in comparison to AIM, PATECTGAN and GEM. On the other hand, Iverson and Terry [27] had both one of the largest domains and the most variables of the papers in our benchmark, as well as a low mutual information between variables. No synthesizer with the exception of GEM exhibited convincing parity performance on this paper.

GEM was the strongest performing synthesizer on one paper (Iverson and Terry [27]). For the other papers in our benchmark, neither PATECTGAN nor GEM were the strongest performing. However, these methods were the most computationally tractable on high-dimensional large-domain data, and were the only methods that were feasible to run on Jeong *et al.* [30], where they both achieved 100% parity. Interestingly, PrivBayes often outperformed MST on our benchmark. We believe that this can be explained by two factors: (1) MST is tailored to work on high-dimensional datasets such as NIST, where explicitly parameterizing a conditional structure (like PrivBayes does) is costly and unstable, while the datasets in our benchmark are relatively low-dimensional; and (2) the findings that comprise the epistemic parity metric are based on conclusions that often rely on conditional relationships, which PrivBayes represents explicitly, while MST does not.

PrivMRF was the slowest synthesizer to run, and required a GPU. This requirement limited our ability to fully assess the capabilities of PrivMRF, although we observe that it performed well on the datasets on which it was able to run successfully. PrivBayes was the second-slowest method to run, due to a known limitation in handling high-dimensional data, but performed competitively on datasets on which it was able to run successfully. Notably, no synthesizer succeeded across all papers, and, remarkably, some findings were *never* reproduced by any of the synthesizers.

Epistemic parity across ϵ values.

Figure 3 compares synthesizer performance across reasonable ϵ values, shown on the x-axis in both sub-figures. The left side of the figure shows aggregated epistemic parity as the percentage of reproduced findings on the y-axis, over all iterations of each synthesizer, averaged over all publications in our benchmark. We observe that synthesizer performance (average parity) improves — although not substantially — for higher ϵ values for marginals-based methods PrivMRF, MST, and AIM. At the smallest values ($\epsilon = e^{-3}, e^{-2}$), the performance of PrivBayes, AIM, and MST all begin to noticeably (and understandably!) degrade, especially on certain findings (e.g., 16-21). Interestingly, PrivBayes achieves best performance at $\epsilon = e$, and PATECTGAN and GEM appear insensitive to the value of ϵ. These trends are consistent with the observations in Figure 2, and support the choice of $\epsilon = e$ as a reasonable privacy budget. Overall, we observe that restricting the privacy budget to $\epsilon = e^{-3}$) does not significantly affect the ability of the synthesizers to reproduce the "easy" findings, while increasing it to $\epsilon = e^2$ does not help with reproducing the "difficult" findings. We conjecture that the modeling structure employed by the synthesizer is more important than the scale of private noise.

The right side of Figure 3 shows average variance of epistemic parity. We observe that variance is lowest for PrivMRF, followed by PrivBayes. Further, we observe that the value of ϵ has little impact on parity variance; AIM is the only synthesizer that benefits from a higher value of ϵ in terms of reduced average parity variance.

The observation that epistemic parity is insensitive to ϵ is significant. It suggests that our metric is substantially different compared to other metrics that were previously used for assessment of DP synthesizers. Parity may provide in-

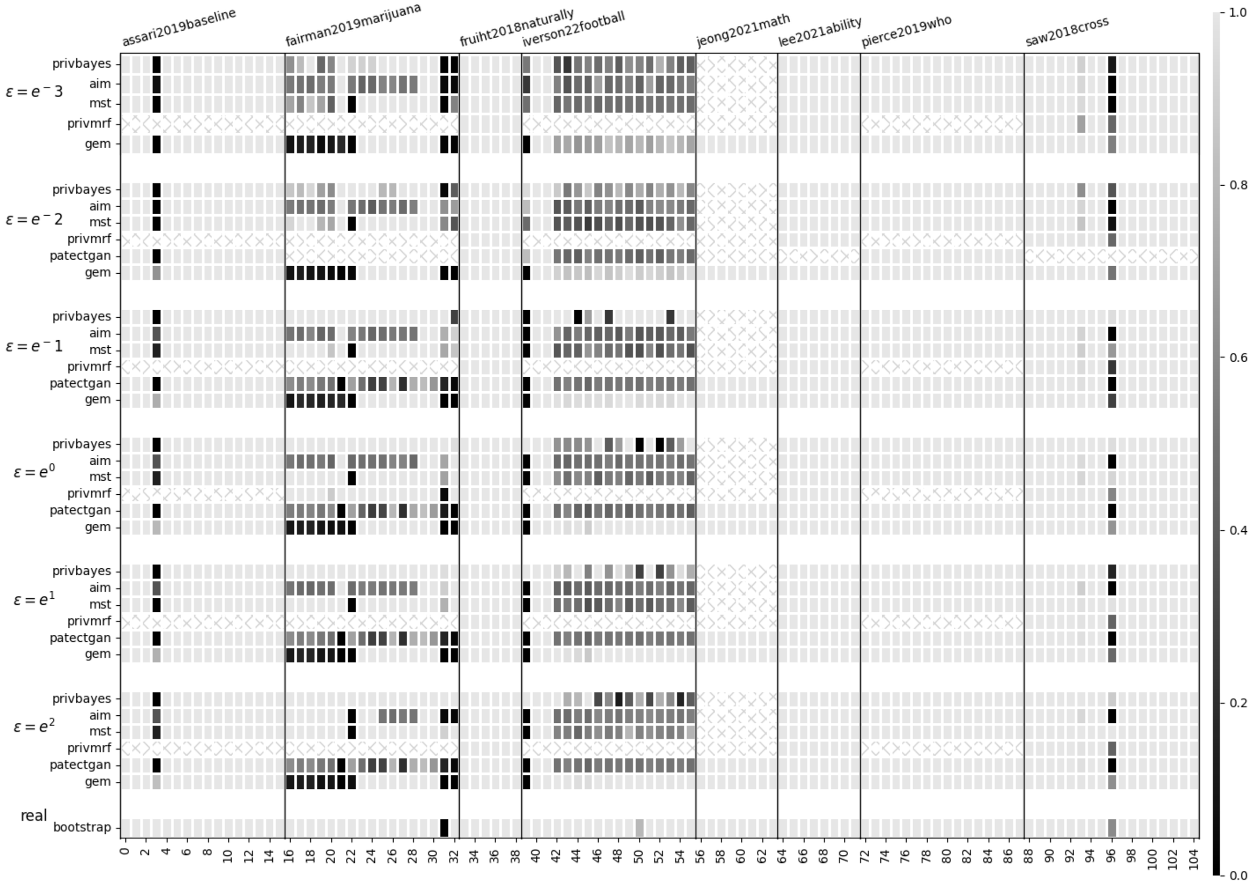

Figure 2: Epistemic parity for six competitive mechanisms for synthesizing data across four ϵ values (e^{-3},e^{-2},e^{-1}, e^0, e^1, e^2). All mechanisms achieve perfect parity on Fruiht and Chan and Lee *et al.*, and all but one achieve perfect parity on Pierce and Quiroz. Only PATECTGAN can scale to support Jeong *et al.*. All methods struggled with the high dimensionality of Iverson and Terry. PrivMRF was too slow to be viable; we report results only for $\epsilon = e^0$. Only PrivBayes and MST achieved reasonable parity for Fairman *et al.*. For datasets associated with Assari and Bazargan and Saw *et al.*, only one finding was difficult to reproduce, and all methods struggled. Surprisingly, parity is relatively insensitive to ϵ.

sight into a more fundamental question about whether a DP synthesizer's *methodology* — the types of measurements it takes to constitute a synthetic distribution — is appropriate to preserve the statistical properties of the dataset that are necessary to reproduce *findings*.

Epistemic parity across finding types.

Table 2 summarizes the methods used in the publications in our benchmark, each corresponding to a type of finding. We observe *Mean Difference* (both *Between-class* and *Temporal*) is by far the most common finding type, followed by *Coefficient Difference*. Whether a finding can be reproduced over DP synthetic data depends on several factors, including dataset size (as in Fairman *et al.* [19]) and dimensionality (as in Iverson and Terry [27]). However, finding type likely also plays a role: The majority (19 out of 26) of *Mean Difference / Temporal* findings are in these two papers that were difficult to reproduce. However, we must be cautious to interpret this as a trend: the remaining 7 findings of type *Mean Difference / Temporal (FC)* were in Saw *et al.* [48], and they were reproduced successfully by all synthesizers. In what follows, we qualitatively evaluate the impact of finding type (and, possibly, of other properties of the finding) on its reproducibility over DP synthetic data.

That some findings are easier to reproduce than others is unsurprising. Though each synthesizer relies on a fundamentally different approach to replicating the joint distribution across all of the data, they each struggle with high dimensional data. Further, for general-purpose synthetic data, PrivMRF, AIM, MST and PrivBayes prioritize lower dimensional 2- or 3-way relationships among variables, and thus it is unsurprising that simple mean comparison findings are easily preserved by these methods.

We were surprised by the high number of findings across all papers (even those that we were unable to replicate) relying only on 1- or 2-dimensional comparisons: The low dimensionality suggests that earlier empirical studies (including Tao *et al.* [51] and Hay *et al.* [24]) may be suitable as proxy tasks. Targeted improvements to the synthesizers may allow us to simultaneously support high utility for individual

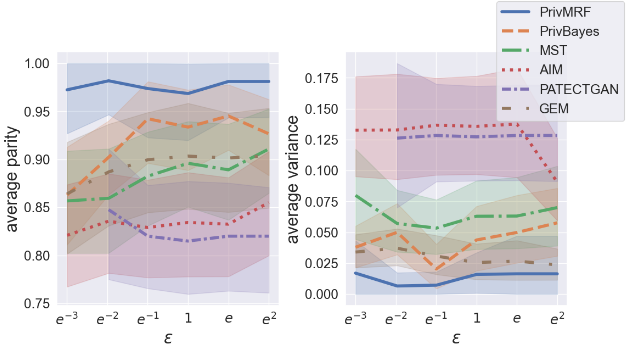

Figure 3: Average epistemic parity across papers achieved by AIM, PrivMRF, MST, PrivBayes, PATECTGAN, and GEM as a function of the privacy parameter $\epsilon \in \{e^{-3}, e^{-2}, e^{-1}, e^{0}, e^{1}, e^{2}\}$. Parity, on the y-axis, is on [0,1] and represents the fraction of reproduced findings over all experiments at each ϵ.

Table 2: Methods used in benchmark papers, each corresponding to a type of *finding* in our framework.

	Descriptive Statistics	8
Regression	Between-Coefficients	4
	Fixed Coefficient (Sign)	2
	Variability	1
Causal Paths	Interaction	1
	Coefficient Difference	19
Logistic Regression	PBR, FNR, FPR	2 (each)
	Accuracy	2
Mean Difference	Between-Class	24
	Temporal (FC)	26
Correlation	Pearson	12
	Spearman	1

findings and their composition into broad conclusions.

Next, we consider 3 findings that were difficult regardless of synthesizer or privacy regimen: #4 (Assari and Bazargan [1]), #39 (Iverson and Terry [27]), and #96 (Saw *et al.* [48]), see Figure 2. Finding #4 is of type *Descriptive Statistics*. It is based on the text statement "Similarly, overall, people had 12.53 years of schooling at baseline (95%CI = 12.34-2.73)." Finding #39 is also of type *Descriptive Statistics*, and is based on a somewhat longer text statement that refers to specific percentages of individuals being diagnosed with specific disorders (5 such pairs of statistics in total). Finding #94 is of type *Mean Difference / Between-class*. It's based on the text statement "From a longitudinal perspective, students from the two lower SES groups—low-middle and low SES groups—had significantly fewer persisters (31.9% and 29.9%) and emergers (6.1% and 5.4%) than their high SES peers (45.1% and 9.0%, respectively)."

These findings were difficult to reproduce because they give specific measurements for variables with large domains. Larger domains require proportionally more DP noise, and so the learned distribution over these variables was too noisy to reproduce the findings within the specified tolerance.

Summary of experimental results.

Overall, we were encouraged by the performance of state-of-the-art synthesizers on our benchmark. DP synthetic data has become more widely used in the social sciences (for Census Data, etc.) and these findings suggest that, in certain contexts, scientists can use DP synthetic data to conduct their scientific inquiry. We caveat this point: *Certain contexts* means relatively low-dimensional tabular data. Our benchmark can be used to assess if those data characteristics hold for a particular dataset, and researchers can proceed with their private analysis with increased confidence.

However, large domain and high-dimensional settings are still a challenge for DP synthesizers: as the domain/number of variables grows, the ease of *fingerprinting* individuals in a dataset increases dramatically. Our findings suggest that existing synthesizers struggle to scale (PrivMRF, MST, AIM, PrivBayes), or are far from achieving reasonable utility (PATECTGAN, GEM). We suggest incorporating more principled methods of data preprocessing, like DP-binning, DP variable pruning, or other domain/variable count reduction techniques into synthesizers, so that successful marginals-based methods can be utilized for more complex data.

6. CONCLUSIONS AND FUTURE WORK

Summary of contributions. We proposed *epistemic parity* as a methodology for measuring the utility of DP synthetic data in support of scientific research. We assembled a benchmark of peer-reviewed papers that analyze one of four studies in the ICPSR social science repository. We then experimentally evaluated epistemic parity achieved by state-of-the-art DP synthesizers over the papers in our benchmark. Overall, we found epistemic parity to be a compelling method for evaluating DP synthesizers. Further, we found that, of the six DP synthesizers we evaluated, no single synthesizer outperformed all others on all papers. Finally, some findings were never reproduced by any of the synthesizers.

Future work: Characterizing false discoveries. Replicating published findings using synthetic versions of the original data can reveal some implications of DP for scientific research. However, this methodology does not assess the possibility of findings that *would have occurred* if the original research had been done on synthetic data, which is related to publication bias [46, 28]. In future work, epistemic parity could be extended to quantify the effect of DP noise in producing these false discoveries by simulating data with both "real" and spurious relationships.

Future work: Rebalancing utility and privacy. Though DP was developed to provide formal guarantees of privacy with best-effort utility, many practitioners and data providers may want the inverse: strong guarantees of utility with quantifiable, flexible risk of privacy violations that can be managed with policy rather than mathematical guarantees. Our benchmark promotes a more holistic discussion of socio-technical-legal systems. Additionally, DP synthesizers can generate arbitrarily large samples at low cost, which makes the power of statistical hypothesis tests another concern for scientific research on private data. Epistemic parity could be extended to estimate the sample size required to achieve a desired power for a particular finding.

7. ACKNOWLEDGEMENTS

This research was supported in part by NSF Awards Nos. 1916505, 1922658, 1934405, and NSF Graduate Research Fellowship Grant No. DGE-2039655, as well as Cisco Award 70618863, the Bill & Melinda Gates Foundation, and the NYU Center for Responsible AI.

8. REFERENCES

[1] S. Assari and M. Bazargan. Baseline obesity increases 25-year risk of mortality due to cerebrovascular disease: role of race. *International Journal of Environmental Research and Public Health*, 16(19):3705, 2019.

[2] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

[3] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 2016.

[4] C. M. Bowen and F. Liu. Comparative study of differentially private data synthesis methods. *Statistical Science*, 35(2):280–307, 2020.

[5] D. Boyd and J. Sarathy. Differential perspectives: Epistemic disconnects surrounding the us census bureau's use of differential privacy. *Harvard Data Science Review (Forthcoming)*, 2022.

[6] M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography: 14th International Conference, TCC 2016-B, Beijing, China, October 31-November 3, 2016, Proceedings, Part I*, pages 635–658. Springer, 2016.

[7] K. Cai, X. Lei, J. Wei, and X. Xiao. Data synthesis via differentially private markov random fields. *Proceedings of the VLDB Endowment*, 14(11):2190–2202, 2021.

[8] M. Christ, S. Radway, and S. M. Bellovin. Differential privacy and swapping: Examining de-identification's impact on minority representation and privacy preservation in the us census. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1564–1564. IEEE Computer Society, 2022.

[9] K. B. Cohen, J. Xia, P. Zweigenbaum, T. J. Callahan, O. Hargraves, F. Goss, N. Ide, A. Névéol, C. Grouin, and L. E. Hunter. Three dimensions of reproducibility in natural language processing. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation*, volume 2018, page 156. NIH Public Access, 2018.

[10] B. Dalton, S. J. Ingels, and L. Fritch. High school longitudinal study of 2009 (hsls:09). 2013 update and high school transcript study: A first look at fall 2009 ninth-graders in 2013. nces 2015-037rev. Technical Report ICPSR36423.v1, Inter-University Consortium for Political and Social Research [distributor], 2016. https://doi.org/10.3886/ICPSR36423.v1.

[11] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. *NeurIPS*, 2021.

[12] J. Ding, X. Zhang, X. Li, J. Wang, R. Yu, and M. Pan. Differentially private and fair classification via calibrated functional mechanism. In *AAAI*, volume 34, pages 622–629, 2020.

[13] I. Dinur and K. Nissim. Revealing information while preserving privacy. In F. Neven, C. Beeri, and T. Milo, editors, *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, pages 202–210. ACM, 2003.

[14] D. Dua and C. Graff. UCI machine learning repository, 2017.

[15] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[16] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 381–390, 2009.

[17] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[18] T. M. Errington, A. Denis, N. Perfito, E. Iorns, and B. A. Nosek. Reproducibility in cancer biology: challenges for assessing replicability in preclinical cancer biology. *Elife*, 10:e67995, 2021.

[19] B. J. Fairman, C. D. Furr-Holden, and R. M. Johnson. When marijuana is used before cigarettes or alcohol: Demographic predictors and associations with heavy use, cannabis use disorder, and other drug-related outcomes. *Prevention Science*, 20(2):225–233, 2019.

[20] V. Fruiht and T. Chan. Naturally Occurring Mentorship in a National Sample of First-Generation College Goers: A Promising Portal for Academic and Developmental Success. 61(3-4):386–397, 2018.

[21] G. Ganev, B. Oprisanu, and E. De Cristofaro. Robin hood and matthew effects–differential privacy has disparate impact on synthetic data. *arXiv preprint arXiv:2109.11429*, 2021.

[22] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. *arXiv preprint arXiv:1012.4763*, 2010.

[23] Harris, Kathleen Mullan and Udry, J. Richard. National longitudinal study of adolescent to adult health (add health), 1994-2018 [public use]. Technical Report ICPSR21600.v25, Inter-university Consortium for Political and Social Research [distributor], Carolina Population Center, University of North Carolina-Chapel Hill [distributor], 2022. https://doi.org/10.3886/ICPSR21600.v25.

[24] M. Hay, A. Machanavajjhala, G. Miklau, Y. Chen, and D. Zhang. Principled evaluation of differentially private algorithms using dpbench. In *Proceedings of the 2016 International Conference on Management of Data*, pages 139–154, 2016.

[25] R. Hill. Evaluating the utility of differential privacy: A use case study of a behavioral science dataset. In *Medical Data Privacy Handbook*, pages 59–82. Springer, 2015.

[26] J. S. House. Americans' changing lives: Waves i, ii, iii, iv, and v, 1986, 1989, 1994, 2002, and 2011. Technical Report ICPSR04690.v9, Inter-university Consortium for Political and Social Research [distributor], 2018. https://doi.org/10.3886/ICPSR04690.v9.

[27] G. L. Iverson and D. P. Terry. High school football and risk for depression and suicidality in adulthood: findings from a national longitudinal study. *Frontiers in neurology*, 12, 2021.

[28] S. Iyengar and J. B. Greenhouse. Selection models and the file drawer problem. *Statistical Science*, pages 109–117, 1988.

[29] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman. Differentially private fair learning. In *ICML*, pages 3000–3008, 2019.

[30] H. Jeong, M. D. Wu, N. Dasgupta, M. Médard, and F. P. Calmon. Whogets the benefit of the doubt? racial bias in machine learning algorithms applied to secondary school math education. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Workshop on Math AI for Education (MATHAI4ED)*, 2021.

[31] D. N. Joanes and C. A. Gill. Comparing measures of sample skewness and kurtosis. *The Statistician*, 47:183–189, 1998.

[32] C. T. Kenny, S. Kuriwaki, C. McCartan, E. T. Rosenman, T. Simko, and K. Imai. The use of differential privacy for census data and its impact on redistricting: The case of the 2020 us census. *Science advances*, 7(41):eabk3283, 2021.

[33] R. Kohavi and B. Becker. UCI adult data set. Technical report, UCI Machine Learning Repository, 1996. https://archive.ics.uci.edu/ml/datasets/adult.

[34] G. Lee and S. D. Simpkins. Ability self-concepts and parental support may protect adolescents when they experience low support from their math teachers. *Journal of Adolescence*, 88:48–57, 2021.

[35] T. Liu, G. Vietri, and S. Z. Wu. Iterative methods for private synthetic data: Unifying framework and new methods. *Advances in Neural Information Processing Systems*, 34:690–702, 2021.

[36] R. McKenna, G. Miklau, M. Hay, and A. Machanavajjhala. Optimizing error of high-dimensional statistical queries under differential privacy. *arXiv preprint arXiv:1808.03537*, 2018.

[37] R. McKenna, G. Miklau, and D. Sheldon. Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *arXiv preprint arXiv:2108.04978*, 2021.

[38] R. McKenna, B. Mullins, D. Sheldon, and G. Miklau. Aim: An adaptive and iterative mechanism for differentially private synthetic data. *arXiv preprint arXiv:2201.12677*, 2022.

[39] National Academies of Sciences, Engineering, and Medicine and others. Reproducibility and replicability in science. 2019.

[40] K. D. R. Pierce and C. S. Quiroz. Who matters most? social support, social strain, and emotions. *Journal of Social and Personal Relationships*, 36(10):3273–3292, 2019.

[41] F. Pinto, C. Soares, and J. Mendes-Moreira. Towards automatic generation of metafeatures. In J. Bailey, L. Khan, T. Washio, G. Dobbie, J. Z. Huang, and R. Wang, editors, *Advances in Knowledge Discovery and Data Mining*, pages 215–226, Cham, 2016. Springer International Publishing.

[42] A. Rivolli, L. P. F. Garcia, C. Soares, J. Vanschoren, and A. C. P. de Leon Ferreira de Carvalho. Characterizing classification datasets: a study of meta-features for meta-learning. *arXiv: Learning*, 2018.

[43] L. Rosenblatt, J. Allen, and J. Stoyanovich. Spending privacy budget fairly and wisely. *CoRR*, abs/2204.12903, 2022.

[44] L. Rosenblatt, B. Herman, A. Holovenko, W. Lee, J. R. Loftus, E. Mckinnie, T. Rumezhak, A. Stadnik, B. Howe, and J. Stoyanovich. Epistemic parity: Reproducibility as an evaluation metric for differential privacy. *Proc. VLDB Endow.*, 16(11):3178–3191, 2023.

[45] L. Rosenblatt, X. Liu, S. Pouyanfar, E. de Leon, A. Desai, and J. Allen. Differentially private synthetic data: Applied evaluations and enhancements. *arXiv preprint arXiv:2011.05537*, 2020.

[46] R. Rosenthal. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638, 1979.

[47] S. Ruggles, C. Fitch, D. Magnuson, and J. Schroeder. Differential privacy and census data: Implications for social and economic research. In *AEA papers and proceedings*, volume 109, pages 403–08, 2019.

[48] G. Saw, C.-N. Chang, and H.-Y. Chan. Cross-sectional and longitudinal disparities in stem career aspirations at the intersection of gender, race/ethnicity, and socioeconomic status. *Educational Researcher*, 47(8):525–531, 2018.

[49] J. Schlimmer. UCI adult data set. Technical report, UCI Machine Learning Repository, 1987. https://archive.ics.uci.edu/ml/datasets/mushroom.

[50] S. Takagi, T. Takahashi, Y. Cao, and M. Yoshikawa. P3gm: Private high-dimensional data release via privacy preserving phased generative model. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 169–180. IEEE, 2021.

[51] Y. Tao, R. McKenna, M. Hay, A. Machanavajjhala, and G. Miklau. Benchmarking differentially private synthetic data generation algorithms. *arXiv preprint arXiv:2112.09238*, 2021.

[52] R. Torkzadehmahani, P. Kairouz, and B. Paten. DP-CGAN: differentially private synthetic data and label generation. *CoRR*, abs/2001.09700, 2020.

[53] United States Department of Health and Human Services. National survey on drug use and health (nsduh), 2014. Technical Report ICPSR36361.v1, Inter-university Consortium for Political and Social Research [distributor], 2016. https://doi.org/10.3886/ICPSR36361.v1.

[54] G. Vietri, G. Tian, M. Bun, T. Steinke, and S. Wu. New oracle-efficient algorithms for private synthetic data release. In *ICML*, pages 9765–9774, 2020.

[55] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[56] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. Privbayes: Private data release via bayesian networks. SIGMOD 2014.