Diverse Community Data for Benchmarking Data Privacy Algorithms

Aniruddha Sen
University of Massachusetts
Amherst, MA 01003

Christine Task
Knexus Research
Oxon Hill, MD 20745
christine.task@knexusresearch.com

Dhruv Kapur
University of Michigan
Ann Arbor, MI 48109

Gary Howarth
NIST
Boulder, CO 80305
gary.howarth@nist.gov
[*]

Karan Bhagat
Knexus Research
Oxon Hill, MD 20745

Abstract

The Diverse Communities Data Excerpts are the core of a National Institute of Standards and Technology (NIST) program to strengthen understanding of tabular data deidentification technologies such as synthetic data. Synthetic data is an ambitious attempt to democratize the benefits of big data; it uses generative models to recreate sensitive personal data with new records for public release. However, it is vulnerable to the same bias and privacy issues that impact other machine learning applications, and can even amplify those issues. When deidentified data distributions introduce bias or artifacts, or leak sensitive information, they propagate these problems to downstream applications. Furthermore, real-world survey conditions such as diverse subpopulations, heterogeneous non-ordinal data spaces, and complex dependencies between features pose specific challenges for synthetic data algorithms. These observations motivate the need for real, diverse, and complex benchmark data to support a robust understanding of algorithm behavior. This paper introduces four contributions: new theoretical work on the relationship between diverse populations and challenges for equitable deidentification; public benchmark data focused on diverse populations and challenging features curated from the American Community Survey; an open source suite of evaluation metrology for deidentified datasets; and an archive of evaluation results on a broad collection of deidentification techniques. The initial set of evaluation results demonstrate the suitability of these tools for investigations in this field.

1 Introduction

Deidentification algorithms take records linked to individuals and attempt to produce data that does not identify individuals but remains useful for analysis. Effective deidentification can allow organizations to share useful information from potentially sensitive data. Such data can be used to train machine learning algorithms; expose fraud, waste, and abuse; improve health outcomes; and other uses. Many mechanisms are available to deidentify data. Some approaches, such as statistical disclosure limitation, redact or suppress information that is deemed particularly identifying. Other approaches, such as differential privacy, and other formal privacy frameworks, develop rigorous

[*] corresponding author

Preprint. Under review.

mathematical definitions of privacy, and algorithms from those frameworks attempt to produce data with quantifiable bounds on how much identifiable information is being shared. Synthetic data algorithms leverage generative models to reproduce sensitive data distributions using new, synthetic individuals. However, while deidentification release mechanisms may (but not necessarily) improve privacy, these mechanisms can also distort data distributions by introducing artifacts and bias. Identifying and resolving these issues is important, but it is not trivial to do. This paper combines new formal contributions, open source tools, and empirical evaluation to push forward progress.

Tools to Benchmark Deidentification Algorithms. To facilitate detailed comparison of deidentification mechanisms, we introduce a set of public, open source tools: the (NIST Diverse Communities Data Excerpts)[1], benchmark data curated from the American Community Survey (ACS); the SDNist Deidentified Data Report tool [2], a suite of data fidelity evaluation methods; and the (NIST Collaborative Research Cycle (CRC) Research Acceleration Bundle), a repository of over 300 deidentification algorithm samples and evaluation results using these tools.

Improving our understanding of data is critical to advance our understanding of algorithm behavior. Our key contribution in this work is the formal analysis of a data property, *subpopulation dispersal*. This concept underlies the tension between diversity, equity, and privacy that impacts all deidentification algorithms operating in real-world conditions. Following this analysis we introduce the Diverse Community Data Excerpts–realistic and tractable benchmark data designed to support rigorous analysis of algorithm behavior under subpopulation dispersal and other challenges. Finally we demonstrate the efficacy of the Excerpts in practice, using visualization metrics from SDNist and synthetic data samples from the CRC project. By iteratively strengthening our understanding of data and algorithms, we believe reliable, equitable and privacy-preserving data release is possible.

2 Distributional Diversity and Subpopulation Dispersal

We use the term *diverse populations* to refer to populations containing two or more subpopulations which differ from each other in terms of their feature correlations. Intuitively, these are more challenging for deidentification algorithms to model accurately. Here, we provide a robust formal analysis of these dynamics which helped inspire the Diverse Community Excerpts benchmark data.

Consider data represented in a histogram, with bins counting the number of occurrences of each record type (more formally defined below). Individuals in bins with small counts are more uniquely identifiable–there are fewer people like them. Many privacy approaches focus on protecting these individuals. Traditional statistical disclosure control techniques like k-anonymity operate by perturbing or redacting records in small bins. Randomization approaches that rely on additive noise, such as differential privacy or subsampling, have much larger relative impact on these small bins. Non-differentially private synthetic data generators may have difficulty modeling sparser areas of the distribution. In general, deidentified data fidelity is more challenging for individuals in small bins (figure 2b). However these individuals are not necessarily represent unimportant outliers. Depending on feature independence it is possible for a large subpopulation to become dispersed into small bins.

When diverse data contains two subpopulations with significantly different patterns of feature correlations, the same schema may disperse one subpopulation and preserve the other, causing the first to be erased by deidentification while the maintaining the second (see figure 1). Diverse real world data, such as the Excerpts, has this property (2a). This is why it is vital to use diverse data to study algorithm behavior. We first introduce two relevant information theoretic tools and our notation. We then formally define *dispersal ratio* and prove bounds on the relationship between diversity and dispersal. We include the proofs as demonstration of theoretical work on data.

Definition 2.1 (Uncertainty Coefficient)**.** The uncertainty coefficient $U(X|F)$ is an information theoretic metric for quantifying correlation between two random variables X and F (in our case an existing feature set F and a new added feature X) [3]. It is a normalized form of mutual information. Note that $0 \leq U(X|F) \leq 1$. It is defined as:

$$U(X|F) = \frac{H(X) - H(X|F)}{H(X)}$$

An uncertainty coefficient of 1 implies that the random variable X can be completely predicted by knowing the value of F, and an uncertainty coefficient of 0 implies that the random variable

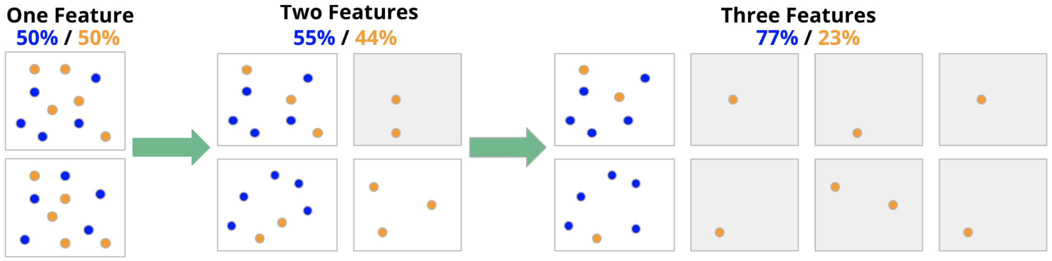
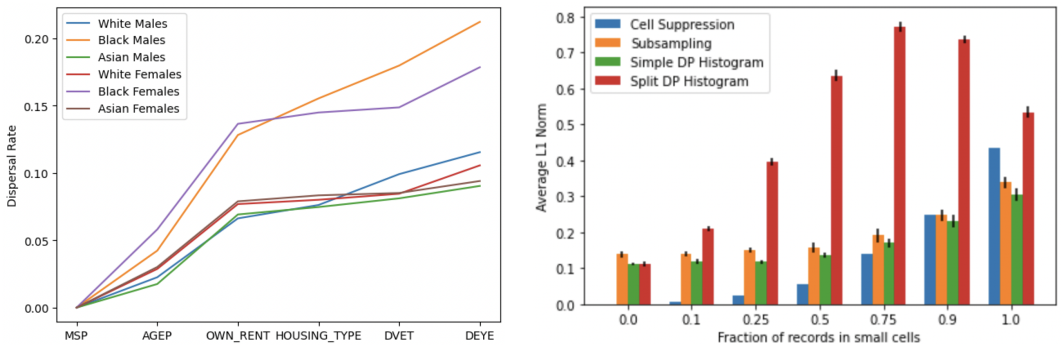


Figure 1: New features are correlated with previous features for members of the blue group, but independent of previous features for members of the orange group. We see that the orange group is increasingly dispersed into small count bins (in grey). These could be erased by deidentification.



(a) Differing dispersal by demographic group in the Diverse Community Excerpts

(b) Increasing error of deidentification by fraction of population dispersed in small count bins

Figure 2: The equity dynamics of diverse data, dispersal, and deidentification

X is independent of F. Thus, $U(X|F)$ inversely correlates with the independence of X for some F. Note that this metric is asymmetric and may not be be equal if the variables are exchanged, i.e. $U(X|F) \neq U(F|X)$. This metric can be used only for categorical data, which makes it useful for a table-based partition schema where the data is distributed into cells with categorical feature labels.

Definition 2.2 (Entropy)**.** The entropy of a discrete random variable X is defined as:

$$\mathrm{H}(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$

where $\mathcal{X}$ denotes the range of X [4]. For this paper, we will be considering the empirical definition of entropy or observed entropy from the probability distribution of a histogram.

Consider a population P with individuals distributed in a table-based partitioned schema S . The probability of an assignment of feature values, corresponding to a bin in the schema S, is calculated as $p(bin_{S(i)}) = \frac{|\{i|bin_{S(i)}\}|}{|P|}$. Let the multivariate distribution of the initial feature set be described by F in the schema S', and the univariate distribution of the new feature be described by X in the schema $(S + X)$. The observed entropy $H(X)$ is dependent on the distribution of X, which remains the same over this operation for any added feature. We assume the size of the population is much larger than the number of bins, i.e. $|P| >> |Range(X, F)|$, to simplify our calculations.

Definition 2.3 (Dispersal Ratio)**.** Let the dispersal ratio for a population P with the addition of feature X be defined as

$$Disperse(S, X, P) = |bin_{(S+X)(P)}|/|bin_{S(P)}|$$

where $bin_{S(P)}$ is defined as the set of all bins in the histogram corresponding to P distributed in the schema S. The main result of this section is proved in Theorem 2.3.

Lemma 2.1. An uncertainty coefficient of 1 is equivalent to a dispersal ratio of 1.

$$U(X|F) = 1 \iff Disperse(S, X, P) = 1$$

Lemma 2.2. An uncertainty coefficient of 0 leads to the maximum dispersal ratio.

$$U(X|F) = 0 \implies \mathit{Disperse}(S, X, P) = |\mathit{Range}(X)|$$

The above lemmas provide good intuition for our main argument in this section. Let's call these the trivial bounds for dispersal ratio on adding a feature X to the schema. Moreover, we are also interested in quantifying how different patterns of feature correlations, represented by the uncertainty coefficient, impact the dispersal ratio for values within these bounds. The following theorem allows us to establish bounds on dispersal ratio as a function of the uncertainty coefficient or independence.

Theorem 2.3. Dispersal Ratio is bounded from above and below as function of the independence of the added feature as follows

$$\frac{|P| \cdot f(u)}{\log(|P|)|\mathit{Range}(F)|} \geq \mathit{Disperse}(S, X, P) \geq \frac{2^{f(u)}}{|\mathit{Range}(F)|}$$

Proof. Lemma 2.1 and Lemma 2.2 give an upper and lower bound for the dispersal ratio which is

$$1 \leq \mathit{Disperse}(S, X, P) \leq |\mathit{Range}(X)|$$

corresponding to

$$1 \geq U(X|F) \geq 0$$

Independence is quantified through the uncertainty coefficient $U(X|F)$. Recall that as U decreases, independence increases, and vice-versa.

Let some arbitrary $u = U(X|F)$. From the definition of $U(X|F)$,

$$u = \frac{H(X) - H(X|F)}{H(X)} \Rightarrow H(X|F) = (1 - u)H(X)$$

Rewriting in terms of the joint entropy [4] $H(X, F)$,

$$H(X, F) = (1 - u)H(X) + H(F) \tag{1}$$

Applying a well-known upper bound on entropy [4], and substituting in equation (1),

$$H(X, F) \leq \log_2(|\mathit{Range}(X, F)|) \implies |\mathit{Range}(X, F)| \geq 2^{(1-u)H(X)+H(F)} \tag{2}$$

Following from our definition of dispersal ratio,

$$\mathit{Disperse}(S, X, P) = \frac{|\mathit{bin}_{(S+X)(P)}|}{|\mathit{bin}_{(S)(P)}|} = \frac{|\mathit{Range}(X, F)|}{|\mathit{Range}(F)|} \tag{3}$$

Thus,

$$\mathit{Disperse}(S, X, P) \geq \frac{2^{(1-u)H(X)+H(F)}}{|\mathit{Range}(F)|} \tag{4}$$

Now, from the definition of entropy,

$$\mathrm{H}(X, F) = -\sum_{x \in (\mathcal{X} \times \mathcal{F})} p(x) \log p(x)$$

Since each bin, corresponding to x in the above equation, must have at least one person in order to contribute to the entropy, $p(x) \geq \frac{1}{|P|}$ where $|P|$ is the size of the population. To analyze any arbitrary distribution, we can first allocate one person to each bin in the range, by definition of range. Now, we have to distribute $|P| - |\mathit{Range}(X, F)|$ people in $|\mathit{Range}(X, F)|$ bins. Entropy is minimized when all the rest of the people are put in one bin. This distribution gives a lower bound for entropy in terms of the range of (X, F). Formally, we get the following inequality,

$$H(X, F) \geq (|\mathit{Range}(X, F)| - 1)\left[\frac{\log(|P|)}{|P|}\right] + \frac{|P| - (|\mathit{Range}(X,F)| - 1)}{|P|} \log\left(\frac{|P|}{|P| - (|\mathit{Range}(X,F)| - 1)}\right) \tag{5}$$

Since $|P| >> |Range(X, F)| >> 1$, the second term is neglected as $1 \cdot \log(\frac{|P|}{|P|}) = 0$,

$$\implies \mathrm{H}(X, F) \geq |Range(X, F)| \left[\frac{\log(|P|)}{|P|} \right]$$

Substituting in equation (1) and rearranging terms, we get

$$Disperse(S, X, P) \leq \frac{|P| \cdot ((1-u)H(X) + H(F))}{\log(|P|)|Range(F)|} \tag{6}$$

We can improve on our trivial bounds of 1 and $|Range(X)|$, from Lemma 2.1 and Lemma 2.2, for the dispersal ratio corresponding to a given u. Combining these two results with the improved upper and lower bounds from equation (3) and equation (5),

$$\min\left\{ \frac{|P|\cdot((1-u)H(X)+H(F))}{\log(|P|)|Range(F)|}, |Range(X)| \right\} \geq Disperse(S, X, P) \geq \max\left\{ \frac{2^{(1-u)H(X)+H(F)}}{|Range(F)|}, 1 \right\} \tag{7}$$

From equation (1), we can write $H(X, F)$ as a function of u i.e. $f(u) := (1-u)H(X) + H(F)$. Also note that as u increases, $f(u)$ decreases and vice-versa. This gives a simpler form for our above equation (also considering only non-trivial bounds),

$$\frac{|P| \cdot f(u)}{\log(|P|)|Range(F)|} \geq Disperse(S, X, P) \geq \frac{2^{f(u)}}{|Range(F)|} \tag{8}$$

□

This result shows that, for some fixed value of entropy of the added feature, the non-trivial upper and lower bounds for the dispersal ratio decrease as the uncertainty coefficient increases, and vice-versa, according to the described behaviour of $f(u)$.

Now, we want to compare the effect of adding a new feature X_1 or X_2 to the schema. Let us assume that X_1 is more "independent" than X_2 of the distribution of P in S (note that this is equivalent to considering a single feature X and two diverse subpopulations $P1, P2$ with differing relationships to X). We can then say that the uncertainty coefficient $u_1 = U(X_1|F)$ is lesser for X_1 as compared to u_2 corresponding to X_2. We can use our results from the above theorem to make this comparison as follows in Theorem 2.4. We define a couple of terms first for ease of notation.

Let the non-trivial lower bound for the dispersal ratio (>1) on adding feature X be denoted as

$$LB(Disperse(S, X, P)) = \frac{2^{(1-u)H(X)+H(F)}}{|Range(F)|} \tag{9}$$

Let the non-trivial upper bound for the dispersal ratio (<$|Range(X)|$) on adding feature X be denoted

$$UB(Disperse(S, X, P)) = \frac{|P| \cdot (1-u)H(X) + H(F)}{\log(|P|)|Range(F)|} \tag{10}$$

Theorem 2.4. Consider two features X_1 and X_2, identical in terms of entropy, that can be added to the schema. If X_1 has higher independence than X_2 with respect to F, it is equivalent to X_1 having a higher LB and higher UB for the dispersal ratio.

$$U(X_1|F) \leq U(X_2|F) \iff LB(Disperse(S, X_1, P) \geq LB(Disperse(S, X_2, P)$$
$$U(X_1|F) \leq U(X_2|F) \iff UB(Disperse(S, X_1, P) \geq UB(Disperse(S, X_2, P)$$

Proof. Consider a population P distributed in a table-based partitioned schema S. Note that higher independence corresponds to a lower uncertainty coefficient. Let $u_1 = U(X_1|F)$, $u_2 = U(X_2|F)$ and $H(X) = H(X_1) = H(X_2)$. The following result can be derived from Theorem 2.3. Let us consider the lower bound first. From equation (9),

$$\frac{LB((Disperse(S, X_1, P))}{LB((Disperse(S, X_2, P))} = \frac{\frac{2^{(1-u_1)H(X)+H(F)}}{|Range(F)|}}{\frac{2^{(1-u_2)H(X)+H(F)}}{|Range(F)|}} \implies \frac{LB((Disperse(S, X_1, P))}{LB((Disperse(S, X_2, P))} = 2^{H(X)(u_2-u_1)}$$

Since entropy is always greater than or equal to 0 , $H(X) \geq 0$ [4]. Thus,

$$LB(Disperse(S, X_1, P) \geq LB(Disperse(S, X_2, P) \iff u_2 - u_1 \geq 0 \iff u_1 \leq u_2$$

Similarly for upper bound, from equation (10), we get

$$\frac{UB((Disperse(S, X_1, P))}{UB((Disperse(S, X_2, P))} = \frac{(1-u_1)H(X) + H(F)}{(1-u_2)H(X) + H(F)}$$

Clearly,

$$UB(Disperse(S, X_1, P) \geq UB(Disperse(S, X_2, P) \iff (1-u_1)H(X)+H(F) \geq (1-u_2)H(X)+H(F)$$
$$\iff u_1 \leq u_2$$

□

3 Introducing the Diverse Community Excerpts

The Excerpts are in the public domain and were designed to explore algorithm behavior on realistic data with diverse subpopulations (see figure 2a). Another key motivation was *tractability*, in light of a recurring problem identified in the NIST Differential Privacy Synthetic Data Challenge [5], NIST Differential Privacy Temporal Map Challenge, and the UNECE High-Level Working Group for the Modernisation of Statistics (HLG-MOS) Synthetic Data Test Drive [6]. Specifically, if the target data is too large or complex, it is very difficult to correctly identify, diagnose and address shortcomings in the deidendified data. This is a serious problem; consumers of deidentified government data cannot afford to overlook even subtle introductions of bias, artifacts, or privacy leaks. But data properties like subpopulation dispersal can induce defects that aren't visible in aggregate utility metrics used by most privacy researchers. And deeper exploration is intractable when considering hundreds of features and millions of records. Addressing these issues requires tools designed to make them accessible.

The Diverse Communities Data Excerpts consist of a small curated geography and feature set derived from the significantly larger 2019 American Community Survey (ACS) Public Use Microdata Sample (PUMS) [7], a publicly available product of the U.S. Census Bureau. The Censu Bureau applies privacy measures as detailed in [8] to their microdata, and no independent privacy risks are created with this publication of this subset of the data. The Excerpts serve as benchmark data for two currently active, open source projects at the National Institute of Standards and Technology– SDNist Deidentified Data Report tool and the 2023 Collaborative Research Cycle (CRC)

The Excerpts' feature set was developed with input from U.S. Census Bureau experts in adaptive sampling design (consider [9]). To identify a small set of communities with challenging, diverse distributions, they leverage previous work on geographical differences in CART-modeled synthetic data (see Appendix B, [10]). The open source SDNist metrics library which accompanies the excerpts was developed with input from HLG-MOS participants and synthetic data contractors working with the U.S. Census. In this section we provide an overview of the Excerpts; in the next section we demonstrate their ability to identify and diagnose problems in deidentification algorithms.

3.1 Data Overview

Feature Selection The original ACS schema contains over four hundred features, which poses difficulties for accurately diagnosing shortcomings in deidentification algorithms. The Excerpts use a small but representative selection of 24 features, covering major census categories: Demographic, Household and Family, Geographic, Financial, Work and Education, Disability, and Survey Weights (which are discussed in the Challenges subsection below). Several Excerpts features were not in the original ACS features, and were designed to provide easier access to certain information. Population DENSITY allows models to distinguish rural and urban geographies, INDP_CAT aggregates industry codes into categories, PINCP_DECILE aggregates incomes into percentile bins relative to the record's state, and EDU simplifies schooling to focus on milestone grades and degrees. The Excerpts usage guidance provides detailed schema information, including several recommended feature subsets for use in diverse subpopulation analyses.

Metadata and Documentation: The Diverse Community Excerpts include JSON data dictionaries with complete feature definitions, a github readme with detailed usage guidance, and illustrative "postcard" introductions to the real-world communities in the data.

Library and Algorithm	Privacy Type	Algorithm Type	Priv. Leak (UEM)	Utility (ES)
DP Histogram (ϵ10)	Differential Privacy (DP)	simple histogram	100%	$\sim$ 90% (988)
R synthpop CART model [11]	non-DP synthetic data	multiple imputation decision tree	2.54%	$\sim$40% (935)
MOSTLY AI SDG [12] [13]	non-DP synthetic data	proprietary pre-trained neural network	0.03%	$\sim$30% (921)
SmartNoise MST (ϵ10) [14]	DP	probabilistic graphical model (PGM)	13.6%	= 10% (969)
SDV CTGAN [15] [16]	non-DP synthetic data	generative adversarial network (GAN)	0.0%	$\sim$5% (775)
SmartNoise PACSynth (ϵ10) [17]	DP + k-anonymity	constraint satisfaction	0.87%	$\sim$1% (551)
synthcity ADSGAN [18] [19]	custom noise injection	GAN	0.0%	$<$ 1% (121)

Table 1: Summary of selected deidentification algorithms. Unique Exact Match (UEM) is a simple privacy metric that counts the percentage of singleton records in the target that are also present in the deidentified data; these uniquely identifiable individuals leaked through the deidentification process. The Equivalent Subsample (ES) utility metric uses an analogy between deidentification error and sampling error to communicate utility; a score of 5% indicates the edit distance between the target and deidentified data distributions is similar to the sampling error induced by randomly discarding 95% of the data. Edit distance is based on the k-marginal metric for sparse distributions. [10], [20]

Challenges: The Excerpts have been designed to be representative of real-world survey data conditions. In addition to diversity, these include other challenging properties: logical constraints between features (e.g. AGE = 6 places a constraint on MARITAL STATUS) that can be difficult for synthetic data generators (see figure 4). Additional modeling challenges include heterogeneous feature types and cardinalities (eg: SEX has two categorical values while INDUSTRY has 271, INCOME is an integer while POVERTY INCOME RATIO is continuous). Uneven feature granularity can amplify problems with unequal subpopulation dispersal (ACS 2019 RAC1P uses one code for Asian, but 4 detailed codes for Native American). We also include the sampling weights that survey data users need to simulate a full population; these lose their original meaning after deidentification changes the data sample, and this is a largely unsolved problem.

Limitations: Although ACS data consumers generally assume the data remains representative of the real population, the ACS PUMS data has had basic statistical disclosure control deidentification applied (as noted above), which may impact its distribution. Additionally, there are shortcomings in the Diverse Community Excerpts that we plan to address in future versions: the Excerpts do not currently contain Household IDs (which would support joining individuals in the same households for social network synthesis), Individual IDs (relevant for reidentification research), or a clear training/testing partition (important for differential privacy research).

4 Demonstration of the Diverse Community Excerpts as Benchmarking Tools

We demonstrate the Diverse Community Excerpts efficacy as a tool for exploring the behavior of deidentification algorithms, using two publicly available resources (section 1): metrics from https://github.com/usnistgov/SDNist [1], and deidentified data samples from the CRC Data and Metrics archive [2]. The archive currently contains over 300 deidentification samples based on the Diverse Community Excerpts target data; we've selected seven interesting cases for this paper (table 1). Full metric documentation, metadata and results are in the Appendix.

The regression metric (figure 3) allows us to explore the impact of variable subpopulation dispersal. Very poor performing algorithms show little impact (ADSGAN, PACSynth), while others have worse performance on the more dispersed subpopulation (black women). CART and MST perform well, but introduce a slight bias that strengthens the correlation between high educational attainment and high income in the dispersed group. This could obscure the real disparity between the groups. Considering privacy, all algorithms in the left column guarantee differential privacy with an identical privacy parameter setting ϵ=10 (weak privacy protection). However we see here and in table 1 that techniques with the same formal guarantee can produce widely varied results in terms of both privacy and utility. PACSynth has additional k-anonymity protection which eliminates rare, dispersed records; this provides good privacy, but has much poorer utility with unusual impacts on the distribution that can be identified in this and other metrics. Meanwhile, the simple DP Histogram provides almost no privacy protection at all, reproducing the target data nearly exactly. MST is a good compromise

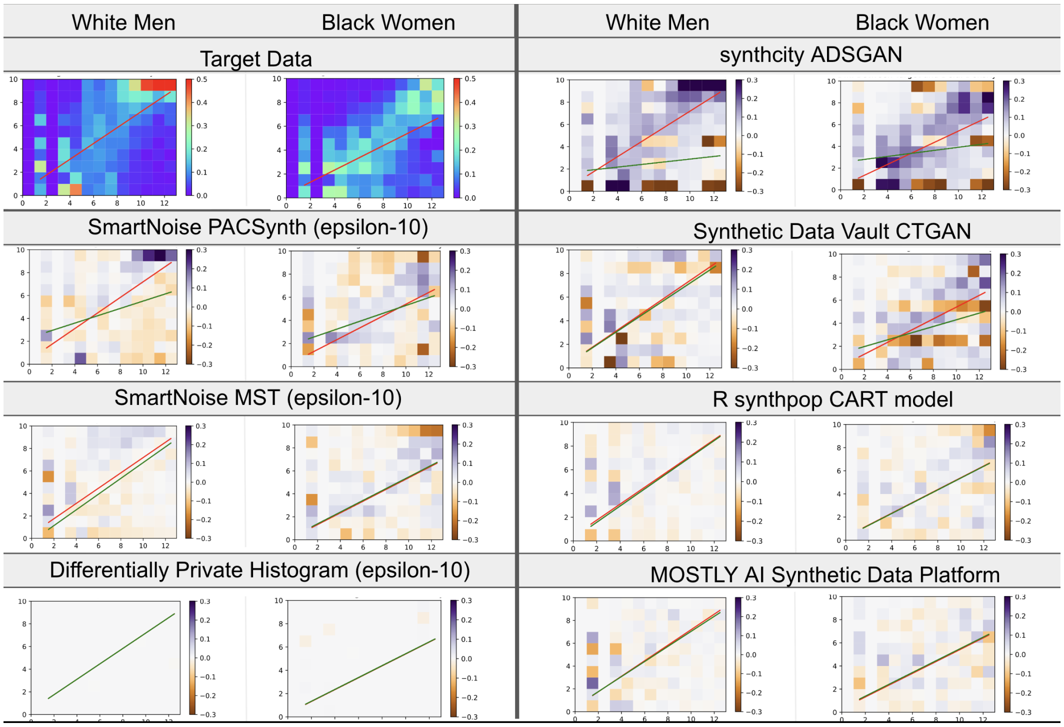


Figure 3: Linear regression metric showing how well the selected algorithms maintain the relationship between educational attainment (x). and income decile (y) for different demographic groups. The target data distribution is shown as a red-blue heatmap of distribution densities (normalized by educational attainment level). The deidentified heatmaps show deviation from the target distribution: brown indicates the deidentified data contains too many individuals in that category, purple indicates it contains too few. The target regression line is given in red, the deidentified line is in green.

of privacy and utility (and would provide better privacy at smaller ϵ), but non-differentially private techniques CART and MostlyAI are better on both privacy and utility.

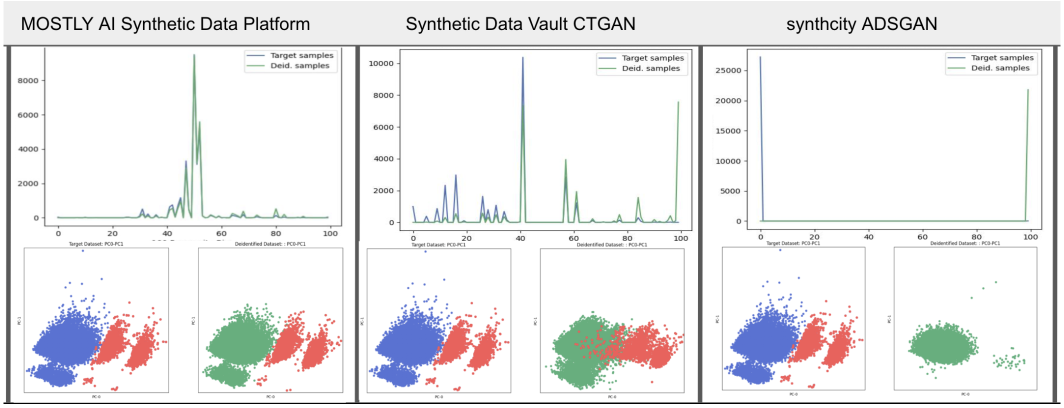


Figure 4: Propensity and PCA metrics. The target data (ground truth) PCA scatterplot is shown on the left in blue, the deidentified data plot is shown on the right, in green; records representing children are highlighted in red. Algorithms that provide good privacy and utility will reproduce the shape of the target scatterplot using new points (ie, synthetic records). The propensity metric uses a classifier to identify synthetic records that clearly defy the target distribution.

In figure 4 we present two more SDNist metrics: The PCA metric uses scatterplots along principle component axes to compare the shapes of distributions. The propensity metric trains a classifier to distinguish between real and synthetic data; if the data are indistinguishable the propensity distributions will peak at the center (indicating the classifier can only make a '50/50 guess' whether a given record is from the target data or deidentified data). If the deidentified data contains artifacts or bias these will be visible as mismatches in the shape of the scatterplot. This figure compares the three neural network synthesizers in our selection, and provides further explanation for the utility values noted in table 1. CTGAN is not preserving constraints that hold for child records (highlighted in red), and is missing a constraint on housing features (visible as the three well-separated clusters in the target data, but blurred together in the CTGAN data). Examples of records violating constraints might be a 7 year old widow or an inmate who owns his jail; these can be confidently classified as synthetic by the propensity metric. Meanwhile, the ADSGAN data contains no children at all, and has retained none of the structure of the target data. Note that these results are more informative than simple edit distance utility scores. We are able to identify and diagnose specific algorithm behaviors induced by the real world challenges in the Excerpt data.

5 Related Works

The Data Responsibly Data Synthesizer, OpenDP, Synthetic Data Vault, Synthcity, NHS Synthetic Data Pipeline, R Rynthpop, and YData libraries have tools to generate and evaluate synthetic data. The SDGym, SDMetrics, Table Evaluator libraries are focused on evaluating synthetic data with arbitrary schema. The Anonymeter and TAPAS tools evaluate privacy of deidentified data.

6 Conclusion

Rather than focus on evaluating arbitrary data, we present tools that are tailored to a specific dataset. We are motivated by previous evaluations of differentially private [21] and of GAN-based synthetic data [22] generators that show surprising differences in algorithm performance. The Excerpts, partnered with SDNist and theory presented here, provide a compact vehicle to address diverse subpopulations, which we show is a persistent problem facing deidentification technologies.

References

[1] Christine Task, Karan Bhagat, Streat Damon, and Gary Howarth. NIST Diverse Community Excerpts Data, December 2022. URL https://data.nist.gov/od/id/mds2-2895.

[2] Christine Task, Karan Bhagat, and Gary Howarth. SDNist v2: Deidentified Data Report Tool, March 2023. URL https://data.nist.gov/od/id/mds2-2943.

[3] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. Numerical Recipes 3rd Edition: The Art of Scientific Computing, 2007.

[4] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[5] Diane Ridgeway, Mary F Theofanos, Terese W Manley, and Christine Task. Challenge design and lessons learned from the 2018 differential privacy challenges. Technical Report NIST TN 2151, National Institute of Standards and Technology (U.S.), Gaithersburg, MD, April 2021. URL https://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.2151.pdf.

[6] United Nations. Economic Commission for Europe et al. Synthetic data for official statistics: a starter guide. 2023. URL https://unece.org/statistics/publications/synthetic-data-official-statistics-starter-guide.

[7] American Community Survey Public Use Microdata Sample (PUMS), 2019. URL https://www.census.gov/programs-surveys/acs/microdata/documentation.2019.html#list-tab-1370939201.

[8] U.S. Census Bureau. Understanding and Using American Community Survey Data: What All Data Users Need to Know. URL https://www.census.gov/programs-surveys/acs/library/handbooks/general.html.

[9] Stephanie Coffey, Benjamin Reist, and Peter V Miller. Interventions On-Call: Dynamic Adaptive Design in the 2015 National Survey of College Graduates. *Journal of Survey Statistics and Methodology*, 8(4): 726–747, September 2020. URL https://academic.oup.com/jssam/article/8/4/726/5533221.

[10] John Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Daniel Kifer, Philip Leclerc, William Sexton, Ashley Simpson, Christine Task, and Pavel Zhuravlev. An Uncertainty Principle is a Price of Privacy-Preserving Microdata, October 2021. URL `http://arxiv.org/abs/2110.13239`.

[11] Beata Nowok, Gillian M. Raab, and Chris Dibben. synthpop: Bespoke creation of synthetic data in r. *Journal of Statistical Software*, 74(11):1–26, 2016. URL `https://www.jstatsoft.org/index.php/jss/article/view/v074i11`.

[12] Mostly.ai. Mostly ai synthetic data platform. `https://mostly.ai/synthetic-data-platform`, 2023.

[13] Michael Platzer and Ivona Krchova. Rule-adhering synthetic data – the lingua franca of learning, 2022. URL `https://arxiv.org/abs/2209.06679`.

[14] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data, August 2021. URL `http://arxiv.org/abs/2108.04978`. arXiv:2108.04978 [cs].

[15] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The Synthetic Data Vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, 2016. URL `https://ieeexplore.ieee.org/document/7796926`.

[16] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling Tabular data using Conditional GAN. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf`.

[17] OpenDP. Private aggregate seeded from pac-synth. `https://docs.smartnoise.org/synth/synthesizers/pac_synth.html`, 2023.

[18] Jinsung Yoon, Lydia N. Drumright, and Mihaela Van Der Schaar. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, August 2020. URL `https://ieeexplore.ieee.org/document/9034117/`.

[19] Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities, 2023. URL `https://arxiv.org/abs/2301.07573`.

[20] Grégorie Lothe, Christine Task, Isaac Slavitt, Nicolas Grislain, Karan Bhagat, and Gary S Howarth. SDNist v1.3: Temporal Map Challenge Environment, December 2021. URL `https://data.nist.gov/od/id/mds2-2515`.

[21] Joshua Snoke, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 181(3):pp. 663–688, 2018. URL `https://www.jstor.org/stable/48547509`. Publisher: [Wiley, Royal Statistical Society].

[22] Yuchao Tao, Ryan McKenna, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Benchmarking Differentially Private Synthetic Data Generation Algorithms, February 2022. URL `http://arxiv.org/abs/2112.09238`.

Checklist

1. For all authors...
   - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   - (b) Did you describe the limitations of your work? [Yes] We address limitations specifically in section 3.1.
   - (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discuss the potential negative social implications of data deidentification in Sections 1, 2 and 3. Examples on specific algorithms are given in Section 4.
   - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   - (a) Did you state the full set of assumptions of all theoretical results? [Yes]

(b) Did you include complete proofs of all theoretical results? [Yes] Main theoretical results have complete proofs in the main paper and additional proofs are in the supplemental information.

3. If you ran experiments (e.g. for benchmarks)...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All code to produce results is public and linked in the paper.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All data and metrics used are public, documented, and linked. Full metadata for the deidentification samples is included in the supplementary material.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See figure 2b. Note that section 5 uses single samples of deidentified data that were contributed to the CRC data archive. All libraries are linked if readers would like to explore variance on a given algorithm.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] All compute time was trivial on typical consumer hardware.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [Yes] Section 3. All included NIST assets are Public Domain: https://www.nist.gov/open/copyright-fair-use-and-licensing-statements-srd-data-software-and-technical-series-publications

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] As a URL

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We use public data published by the U.S. Census Bureau, and in the text of the paper we link to the Bureau's documentation on data collection and privacy protections.

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We use public data published by the U.S. Census Bureau, and in the text of the paper we link to the Bureau's documentation on data collection and privacy protections. We do not identify any offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We collected no original data.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We collected no original data.

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We collected no original data.

Supplemental Information for "Diverse Community Data for Benchmarking Data Privacy Algorithms"

June 19, 2023

SI Contents

A Dataset Provisions

A.1 Dataset URLs

- The NIST public data repository access point.
- Direct data access link.
- Direct data access to the data dictionary.
- The dataset DOI is 10.18434/mds2-2895.

A.1.1 Dataset Format Notes

The raw data are in CSV format with JSON data dictionaries defining valid values.

The NIST data repository has a structured metadata retrieval system that interfaces with data.gov and conforms to FAIR principles and the best practice for Federal Data Strategy. See additional information here.

A.2 Author Statement

The authors bear all responsibility in case of violation of rights. We have confirmed licensing and provide detailed information in the article and in the dataset datasheet.

A.3 Hosting and Licensing

The data associated with this publication were created, hosted, and maintained, by the National Institute of Standards and Technology in their permanent data repository, in perpetuity.

The data are in the public domain. NIST statement on software and data:
"NIST-developed software is provided by NIST as a public service. You may use, copy, and distribute copies of the software in any medium, provided that you keep intact this entire notice. You may improve, modify, and create derivative works of the software or any portion of the software, and you may copy and distribute such modifications or works. Modified works should carry a notice stating that you changed the software and should note the date and nature of any such change. Please explicitly acknowledge the National Institute of Standards and Technology as the source of the software. NIST-developed software is expressly provided "AS IS." NIST MAKES NO WARRANTY OF ANY KIND, EXPRESS, IMPLIED, IN FACT, OR ARISING BY OPERATION OF LAW, INCLUDING, WITHOUT LIMITATION, THE IMPLIED WARRANTY OF

B Datasheet for dataset “NIST Diverse Communities Data Excerpts”

Questions from the Datasheets for Datasets paper, v7.

Jump to section:

- Motivation
- Composition
- Collection process
- Preprocessing/cleaning/labeling
- Uses
- Distribution
- Maintenance

B.1 Motivation

B.1.1 For what purpose was the dataset created?

The NIST Diverse Communities Data Excerpts (the Excerpts) are demographic data created as benchmark data for deidentification technologies.

The Excerpts are designed to contain sufficient complexity to be challenging to de-identify and with a compact feature set to make them tractable for analysis. We also demonstrate the data contain subpopulations with varying levels of feature independence, which leads to small cell counts, a particularly challenging deidentification problem.

The Excerpts serve as benchmark data for two open source projects at the National Institute of Standards and Technology (NIST): the SDNist Deidentified Data Report tool and the 2023 Collaborative Research Cycle (CRC).

B.1.2 Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The Excerpts were created by the Privacy Engineering Program of the Information Technology Laboratory at the National Institute of Standards and Technology (NIST).

The underlying data was published by the U.S. Census Bureau as part of the 2019 American Community Survey (ACS) Public Use Microdata Sample (PUMS).

B.1.3 Who funded the creation of the dataset?

The data were collected by the U.S. Census Bureau, and the Excerpts were curated by NIST. Both are U.S. Government agencies within the Department of Commerce. Aspects of the Excerpts creation were supported under NIST contract 1333ND18DNB630011.

B.1.4 Any other comments?

No.

B.2 Composition

B.2.1 What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The instances in the data represent individual people.

The Excerpts consist of a small curated geography and feature set derived from the significantly larger 2019 American Community Survey (ACS) Public Use Microdata Sample (PUMS), a publicly available product of the U.S. Census Bureau. The original ACS schema contains over four hundred features, which poses difficulties for accurately diagnosing shortcomings in deidentification algorithms. The Excerpts use a small but representative selection of 24 features, covering major census categories: Demographic, Household and Family, Geographic, Financial, Work and Education, Disability, and Survey Weights. Several Excerpts features are derivatives of the original ACS features, designed to provide easier access to certain information (such as income decile or population density).

There is only one type of instance. All records in the data represent separate, individual people.

B.2.2 How many instances are there in total (of each type, if appropriate)?

There are three geographic partitions in the data. See the "postcards" and data dictionaries in each respective directory for more detailed information. Instances in partitions:

- `national`: 27254 records
- `massachusetts`: 7634 records
- `texas`: 9276 records

B.2.3 Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The data set is a curated sample of the ACS by geography, with a reduced feature set designed to provide a tractable foundation for benchmarking deidentification algorithms (24 features rather than the original ACS's 400 features). Geographically it is comprised of 31 Public Use Microdata Areas

- `national`: 27254 records drawn from 20 Public Use Microdata Areas (PUMAs) from across the United States. This excerpt was selected to include communities with very diverse subpopulation distributions.
- `texas`: 9276 records drawn from six PUMAs of communities surrounding Dallas-Fort Worth, Texas area. This excerpt was selected to focus on areas with moderate diversity.
- `massachusetts`: 7634 records drawn from five PUMAs of communities from the North Shore to the west of the greater Boston, Massachusetts area. This excerpt was selected to focus on areas with less diversity.

B.2.4 What data does each instance consist of?

The instances are individual, tabular data records in CSV format with Demographic, Household and Family, Geographic, Financial, Work and Education, Disability, and Survey Weights features.

In addition, there is metadata and documentation: schema files for each of the three geographies containing the features and valid data ranges in JSON format, and 'postcard' documentation with English-language descriptions of the areas described by the data in PDF format. There is also an overarching readme and data dictionary.

B.2.5 Is there a label or target associated with each instance?

No. These data are not designed specifically for classifier tasks.

B.2.6 Is any information missing from individual instances?

There is no missing information in these excerpts, all records are complete.

B.2.7 Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Relationships between records have not been included in this version of the data. Although the Excerpts data does contain multiple individuals from the same household, it does not include the ACS PUMS Household ID or relationship features needed to join them into a network. We expect to include those features in a future update to the Excerpts.

B.2.8 Are there recommended data splits (e.g., training, development/validation, testing)?

There are three geographic partitions to facilitate benchmarking algorithms on populations with differing levels of heterogeneity/diversity (MA, TX and National). There are no splits designed specifically for training and testing purposes. All of the data presented at this time are from the 2019 ACS collection. In the future we plan to add additional years.

B.2.9 Are there any errors, sources of noise, or redundancies in the dataset?

Although ACS data consumers generally assume the data remains representative of the real population, the ACS PUMS data has had basic statistical disclosure control deidentification applied (including swapping and subsampling), which may impact its distribution. For more information, see documentation from the U.S. Census Bureau.

B.2.10 Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

All of the data is self-contained within the repository. The data are drawn from public domain sources and thus have no restrictions on usage.

B.2.11 Does the dataset contain data that might be considered confidential?

The Excerpts are a subset public data published by the U.S. Census Bureau. The U.S. Census Bureau is bound by law, under Title 13 of the U.S. Code, to protect the identities of individuals represented by the data. See here for details on the Census' data stewardship. The Census takes elaborate steps to reduce risk of re-identification of individuals surveyed and provide information regarding their suppression scheme here.

B.2.12 Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

B.2.13 Does the dataset relate to people?

Yes.

B.2.14 Does the dataset identify any subpopulations (e.g., by age, gender)?

The data includes demographic features such as Age, Sex, Race and Hispanic Origin which may be used to disaggregate by subpopulation. It additionally includes non-demographic features such as Educational Attainment, Income Decile and Industry Category which also produce subpopulation distributions with disparate patterns of feature correlations.

The racial and ethnicity subpopulation breakdown by geography is as follows (note that hispanic origin and race are separate features):

- `MA Dataset (less diverse)`: 4% Hispanic and 89% White, 7% Asian, 2% Black, 2% Other, 0% AIANNH
- `TX Dataset (more diverse)`: 19% Hispanic and 85% White, 7% Black, 4% Other, 3% Asian, 1% AIANNH
- `National Dataset (especially diverse)`: 10% Hispanic and 56% White, 22% Black, 10% Other, 9% Asian, 3% AIANNH

B.2.15 Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

The Excerpts are survey results from real individuals as collected by the U.S. Census Bureau. See response above for more information about Census' data protections.

The subset of the Census' data that we provide here introduces no additional information, and therefore does not increase the risk of identifying individuals.

B.2.16 Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

Yes. These data are detailed demographic records. See response above for more information about Census Bureau's data protections.

B.2.17 Any other comments?

No.

B.3 Collection process

B.3.1 How was the data associated with each instance acquired?

This data is a curated geographic subsample of the 2019 American Community Survey Public Use Microdata files. The U.S. Census Bureau details its survey data collection approach here.

B.3.2 What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

See previous response.

B.3.3 If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The data set is a (deterministic) curated sample by geography. It is comprised of 31 Public Use Microdata Areas

- `national`: 27254 records drawn from 20 Public Use Microdata Areas (PUMAs) from across the United States. This excerpt was selected to include communities with very diverse subpopulation distributions.
- `texas`: 9276 records drawn from six PUMAs of communities surrounding Dallas-Fort Worth, Texas area. This excerpt was selected to focus on areas with moderate diversity.
- `massachusetts`: 7634 records drawn from five PUMAs of communities from the North Shore to the west of the greater Boston, Massachusetts area. This excerpt was selected to focus on areas with less diversity.

B.3.4 Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

[See response above.](### How was the data associated with each instance acquired?

B.3.5 Over what timeframe was the data collected?

This data was collected during 2019.

B.3.6 Were any ethical review processes conducted (e.g., by an institutional review board)?

The Excerpts are a curated subsample of existing public data published by the U.S. Government. No IRB review was necessary by institution policy.

B.3.7 Does the dataset relate to people?

Yes.

B.3.8 Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Other Sources. This data is a curated geographic subsample of the 2019 American Community Survey Public Use Microdata files, which are available here.

B.3.9 Were the individuals in question notified about the data collection?

Yes. See response above.

B.3.10 Did the individuals in question consent to the collection and use of their data?

Yes. See response above.

B.3.11 If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

See response above.

B.3.12 Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

This data is a curated geographic subsample of the 2019 American Community Survey (ACS) Public Use Microdata files. Many investigations have examined ACS data with some information published by the Census Bureau itself.

The data presented here, the Excerpts, are a subset of the data and present no additional risks to the subjects surveyed by the Census.

B.3.13 Any other comments?

No.

B.4 Preprocessing/cleaning/labeling

B.4.1 Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

The original ACS data is clean and no class labeling was done. However, several Excerpts features are new derivatives of ACS features designed to provide easier access to certain information. Population DENSITY divides PUMA population by surface area and allows models to distinguish rural and urban geographies. INDP_CAT aggregates detailed industry codes into a small set of broad categories. PINCP_DECILE aggregates incomes into percentile bins relative to the record's state. And, EDU simplifies the original ACS schooling feature to focus on milestone grades and degrees.

B.4.2 Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

See the U.S. Census Bureau's documentation for information about published ACS data.

B.4.3 Is the software used to preprocess/clean/label the instances available?

The preprocessing was minimal (addition of a small set of derivative features), and can be reproduced as described above. The code is not currently available.

B.4.4 Any other comments?

No.

B.5 Uses

B.5.1 Has the dataset been used for any tasks already?

The Excerpts serve as benchmark data for two open source projects at the National Institute of Standards and Technology (NIST): the SDNist Deidentified Data Report tool and the 2023 Collaborative Research Cycle (CRC).

B.5.2 Is there a repository that links to any or all papers or systems that use the dataset?

No. Users are not mandated to contribute their work to any central repository. We publish user-contributed data here. We recommend that data users cite our work using the dataset DOI.

B.5.3 What (other) tasks could the dataset be used for?

The Excerpts were designed for benchmarking privacy-preserving data deidentification techniques such as synthetic data or statistical disclosure limitation. However, they can be used to study the behavior of any tabular data machine learning or analysis technique when applied to diverse populations. Synthetic data generators are just an especially verbose application of machine learning (producing full records rather than class labels), so tools designed to improve understanding of synthetic data have potential for a much broader application.

B.5.4 Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

The U.S. Census Bureau recommends using sampling weights to account for survey undersampling and generate equitable full population statistics. The PWGPT feature included in the Excerpts is the person (record) level sampling weight. For full population statistics, each record should be multiplied by its sampling weight.

B.5.5 Are there tasks for which the dataset should not be used?

The Excerpts are suitable for any application relevant to government survey data over the selected feature set.

B.5.6 Any other comments?

No.

B.6 Distribution

B.6.1 Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

B.6.2 Has the dataset been used for any tasks already?

The Excerpts serve as benchmark data for two open source projects at the National Institute of Standards and Technology (NIST): the SDNist Deidentified Data Report tool and the 2023 Collaborative Research Cycle (CRC).

B.6.3 How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

10.18434/mds2-289

B.6.4 When will the dataset be distributed?

The dataset is currently available to the public.

B.6.5 Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The data are in the public domain. See the following statement from NIST.

B.6.6 Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No. All data are drawn from public domain sources.

B.6.7 Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No. All data are drawn from public domain sources and have no known export or regulatory restrictions.

B.6.8 Any other comments?

No.

B.7 Maintenance

B.7.1 Who is supporting/hosting/maintaining the dataset?

This dataset is hosted by NIST and maintained by the Privacy Engineering Program.

B.7.2 How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Dataset managers can be reached by raising an issue, emailing the Privacy Engineering Program, or by contacting the project principal investigator, Gary Howarth.

B.7.3 Is there an erratum?

There have been small updates to the meta-data data dictionary.json files (for example, to improve clarity in descriptive strings for features). The data are maintained in a public GIt repository and thus all changes to the data are recorded in a public ledger.

B.7.4 Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Since the data are excerpts from the 2019 release of the American Community Survey, we do not expect any updates to labels or instances. We do plan on one mayor updated version release in the future with the following improvements:

- `Household ID features`: Allows joins between individuals in the same household
- `Individual ID`: Supports reidentification research.
- `Training Data Partition`: Including excerpts from 2018 for algorithm development/training and as a baseline for reidentification studies
- `Large-sized low-diversity excerpt`: Our current low-diversity excerpts, MA and TX, have much fewer records than our high-diversity excerpt, National; this can be a confounding factor for comparative analyses.

B.7.5 If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

These data are in the public domain and as such there are no retention limits.

B.7.6 Will older versions of the dataset continue to be supported/hosted/maintained?

The data are maintained in a public Git repository and thus all changes to the data are recorded in a public ledger. There are specific releases in the repository that capture major data milestones.

B.7.7 If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

We invite the public to use and build on these resources. First, these resources are provided by NIST as a public service, and the public is free to integrate these resources into their own work. Second, we invite the public to raise issues in the dataset repository, allowing for a transparent interaction. Individuals and groups wishing to make substantial contributions are encouraged to contact the project principal investigator, Gary Howarth.

B.7.8 Any other comments?

No.

C Math Appendix

C.1 Proofs of Lemmas 2.1 and 2.2 (and additional material)

We introduced the concept of dispersal ratio in the main paper with the purpose of a giving the reader a clear and intuitive explanation of the term. In doing so, we omitted some formal results that might be interesting to examine in order to understand the mechanics behind dispersal ratio and independence. The perceptive reader may have noticed that we stated two lemmas in section 2 without proving them. Recall the definition of dispersal ratio.

Definition C.1 (Dispersal Ratio)**.** Let the dispersal ratio for a population P with the addition of feature X be defined as

$$Disperse(S, X, P) = |bin_{(S+X)(P)}|/|bin_{S(P)}|$$

We begin by providing proofs of Lemma 2.1 (corresponding to Lemma C.1) and Lemma 2.2 (corresponding to Lemma C.2) as stated in section 2. We follow it up with a result that may be of interest. These proofs follows the same framework and terminology as used in the main paper.

Lemma C.1. An uncertainty coefficient of 1 is equivalent to a dispersal ratio of 1.

$$U(X|F) = 1 \iff Disperse(S, X, P) = 1$$

Proof.

$$U(X|F) = 1 \Rightarrow \frac{H(X) - H(X|F)}{H(X)} = 1 \Rightarrow H(X|F) = 0$$

Consider the following result. $H(X|F) = 0$ if and only if X is a function of F i.e. $\forall f : p(f) > 0$, there is only one possible value of x with $p(x, f) > 0$ [1].

Let the function g between X and F be denoted by $F = g(X)$. Applying the result here, let there be m elements in the domain of F, which implies there can be no more than m elements in the co-domain of X, to constitute a valid function. Let the elements in the range of F be denoted by $f_1, f_2 ... f_m$, and that of X be denoted by $x_1, x_2 ... x_{m'}$.
Since X is a function of F, there exists only one element $x_i \in X$ corresponding to $f_j \in F$.
Thus, all bins in the schema $(S + x)$ can be denoted by $(f_i, x_i) = (f_i, g(f_i))$. Since there are m bins, corresponding the size of the domain, in F, there will be exactly m bins in $F' = (F, X)$. Therefore,

$$|bin_{(S+X)(P)}| = |bin_{S(P)}|$$

$$\Rightarrow Disperse(S, X, P) = 1$$

Similarly, the converse of the lemma can be proved by taking the converse of the above result and considering that the inverse of the function $g' = g^{-1}(x)$ for x: $(F, X) \rightarrow F$ is uniquely defined if the dispersal ratio is 1. □

Lemma C.2. An uncertainty coefficient of 0 leads to the maximum dispersal ratio.

$$U(X|F) = 0 \implies Disperse(S, X, P) = |Range(X)|$$

Proof.

$$U(X|F) = 0 \implies \frac{H(X) - H(X|F)}{H(X)} = 0 \implies H(X|F) = H(X)$$

This implies X and F are independent observations [1]. Observe that the range of $Y = (X, F)$ can take maximum $n_{max} = |Range(X)||Range(F)|$ values since that it is the number of elements in $X \times F$. Note that $|Range(F)| = |bin_{S(P)}|$. Here, $Range(Y) = n_{max}$ due to the independence of X and F since

$$\forall x, f : Pr[Y = y] = (Pr[X = x] * Pr[F = f]) \neq 0$$

As there are n_{max} non-zero values for the probability distribution of Y, the size of the range of Y is maximum. Note that Y exactly expresses the distribution of values in the schema $(S + X)$.
Therefore,

$$\begin{aligned} &|bin_{(S+X)(P)}| = |bin_{S(P)}||Range(X)| \\ \implies &Disperse(S, X, P) = |Range(X)| \end{aligned}$$

which is the maximum dispersal ratio since $|bin_{S(P)}| * |Range(X)|$ was maximized. □

We now show an interesting consequence of the relation between dispersal ratio and the initial population. The following lemmas prove that a small population size can lead to small cell counts.

Consider a population P with a sub-population P_1, distributed in a table-based partitioned schema. Consider an individual $i \in P_1$, who gets placed in a bin under schema S. We denote the size of that bin as $size(bin_{S(i)})$. Let a feature f be added to the schema.

Lemma C.3. The dispersal ratio is always greater than or equal to 1.

$$Disperse(S, f, P_1) \geq 1$$

Proof. Consider an arbitrary $bin_S(i)$ in the schema S with the m features in the feature set $f_1, f_2, f_3 ... f_m$.

Adding a new feature f to the schema S with feature values (say) in the set $V = \{v_1, v_2\}$ will subdivide all records in $f_1, f_2, f_3 ... f_m$ into $f_1, f_2, f_3 ... f_m, v_1$ and $f_1, f_2, f_3 ... f_m, v_2$, by the definition of partitioning.

$bin_S(i)$ in the schema S will be replaced by at least one bin or more, in the schema $(S+f)$. Thus, the dispersal ratio for the sub-population of $bin_{S(i)} : i \in P_1$ is always greater than 1.

Since for each disjoint sub-population corresponding to each bin $\in S$, this ratio is greater than one, the dispersal ratio for the overall population P_1 over the schema S and adding a new feature f, is also greater than 1. □

Definition C.2 (Average bin size for population P_1)**.** It is defined as

$$\left[\sum_{S(i):i \in P_1} size(bin_{S(i)}) \right] / |bin_{S(P_1)}|$$

Lemma C.4. If a new feature f is added to the schema denoted by $S + f$, then the average bin size will stay the same or decrease.

$$\left[\sum_{S(i):i\in P_1} size(bin_{S(i)})\right] /|bin_{S(P_1)}| \geq \left[\sum_{(S+f)(i):i\in P_1} size(bin_{(S+f)(i)})\right] /|bin_{(S+f)(P_1)}|$$

Proof. For each of the disjoint partitions of some $S(i) : i \in P_1$, records of the form $i \in P_1$ do not get merged with any records that were not in the initial bin $S(i)$, by definition of partitioning. Thus, summing over all such bins,

$$\left[\sum_{S(i):i\in P_1} size(bin_{S(i)})\right] \geq \left[\sum_{(S+f)(i):i\in P_1} size(bin_{(S+f)(i)})\right] \tag{1}$$

Note that there is a '$\geq$' inequality since there may be bins in the schema $S + f$ that do contain records of the form $i \in P_1$, which were previously grouped with records $i \in P_1$ in the schema S. From Lemma C.3, if the dispersal ratio for population P_1 is r_1, then $r_1 \geq 1$, which implies

$$|bin_{S(P_1)}| \leq |bin_{(S+f)(P_1)}| \tag{2}$$

Combining equations (1) and (2) proves our result, by observing that they are the numerator and denominator respectively of our desired inequality. □

Assume two sub-populations P_0 and P_1 are distributed in the same arbitrary number of bins $|bin_{S(P_1)}| = |bin_{S(P_0)}| = m$. If on adding a feature f, P_0 and P_1 have the same dispersal ratio ($r_0 = r_1 = r'$), then $|bin_{(S+f)(P_1)}| = |bin_{(S+f)(P_0)}| = mr'$. The ratio of their average bin sizes for the schema $S + f$ is

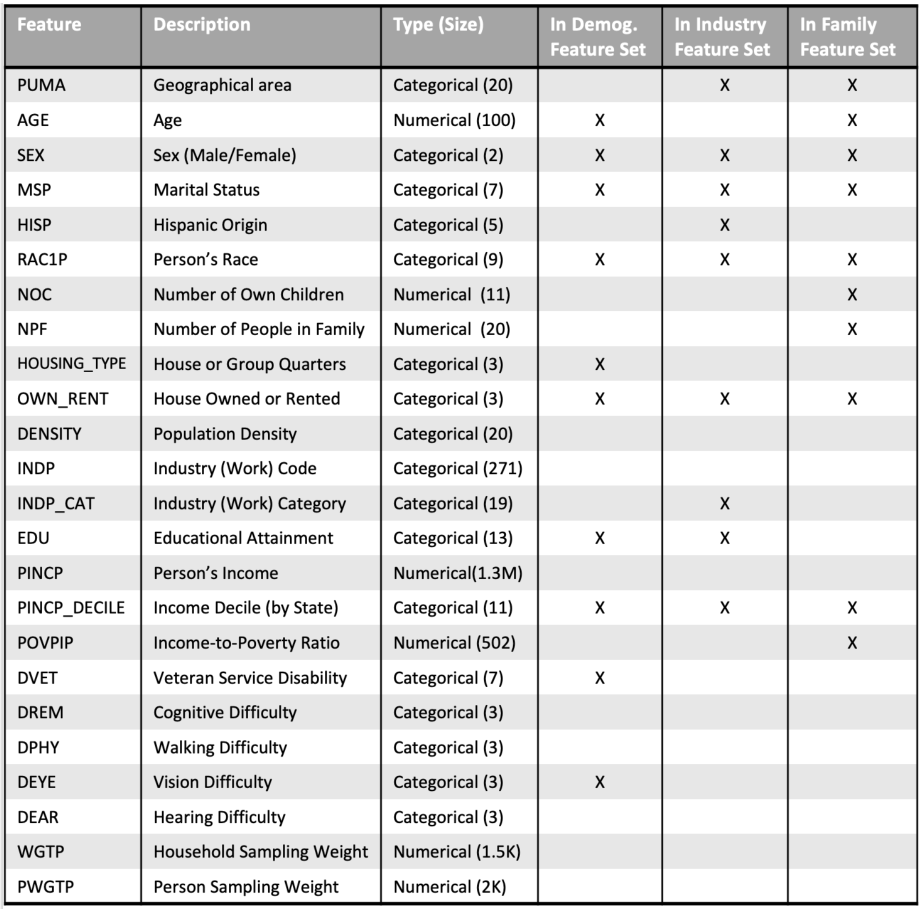
$$\frac{\frac{\left[\sum_{S+f} size(bin_{(S+f)(i)})_{P_1}\right]}{mr'}}{\frac{\left[\sum_{S+f} size(bin_{(S+f)(i)})_{P_0}\right]}{mr'}}$$

The average bin size is directly correlated to the size of the sub-population for the same initial number of bins and the same dispersal ratio. Therefore, if one subgroup (say P_0) is smaller than the other (P_1), then the average bin size for P_0 is less than that of P_1.

As the average bin size drops for members of a sub-population, the utility will also drop monotonically for partition-based algorithms.

D Feature Definitions and Recommended Subsets

Figure 1 lists the 24 Excerpts features. The majority are from the 2019 American Community Survey Public Use Micodata; four of them (DENSITY, INDP_CAT, EDU, PINCP_DECILE) were derived from ACS features or public data as described in B.4. Along with feature type, we've included cardinality (number of possible values). Because some deidentification algorithms require small feature spaces, the NIST CRC program recommends three smaller feature subsets: Demographic-focused, Industry-focused and Family-focused. Each subset showcases different feature mechanics, while sharing common features to delineate subpopulations (SEX, MSP, RAC1P, OWN_RENT, PINCP_DECILE).



Feature	Description	Type (Size)	In Demog. Feature Set	In Industry Feature Set	In Family Feature Set
PUMA	Geographical area	Categorical (20)		X	X
AGE	Age	Numerical (100)	X		X
SEX	Sex (Male/Female)	Categorical (2)	X	X	X
MSP	Marital Status	Categorical (7)	X	X	X
HISP	Hispanic Origin	Categorical (5)		X	
RAC1P	Person's Race	Categorical (9)	X	X	X
NOC	Number of Own Children	Numerical (11)			X
NPF	Number of People in Family	Numerical (20)			X
HOUSING_TYPE	House or Group Quarters	Categorical (3)	X		
OWN_RENT	House Owned or Rented	Categorical (3)	X	X	X
DENSITY	Population Density	Categorical (20)			
INDP	Industry (Work) Code	Categorical (271)			
INDP_CAT	Industry (Work) Category	Categorical (19)		X	
EDU	Educational Attainment	Categorical (13)	X	X	
PINCP	Person's Income	Numerical(1.3M)			
PINCP_DECILE	Income Decile (by State)	Categorical (11)	X	X	X
POVPIP	Income-to-Poverty Ratio	Numerical (502)			X
DVET	Veteran Service Disability	Categorical (7)	X		
DREM	Cognitive Difficulty	Categorical (3)			
DPHY	Walking Difficulty	Categorical (3)			
DEYE	Vision Difficulty	Categorical (3)	X		
DEAR	Hearing Difficulty	Categorical (3)			
WGTP	Household Sampling Weight	Numerical (1.5K)			
PWGTP	Person Sampling Weight	Numerical (2K)			

Figure 1: The 24 Features in the Excerpts, and recommended feature subsets.

E Detailed Evaluation Reports and Metadata on Selected Deidentified Data Samples

As we noted in the main paper, the NIST CRC Data and Metrics Bundle is an archive of 300 deidentifed data samples and evaluation metric results. To demonstrate the efficacy of the Excerpts for identifying and diagnosing behaviors of deidentificaiton algorithms on diverse populations, we selected seven algorithms from the archive to showcase in the paper. Below we provide the complete metadata and highlighted PCA plot for each sample, as well as links to their detailed evaluation reports (available online in the sample report section of the SDNist repository).

Each detailed evaluation report contains the metrics listed below, along with complete results, detailed metric definitions accessible to non-technical stakeholders, a human-readable data dictionary, and additional references.

SDNist Detailed Report Metrics List:

- K-marginal Edit Distance
- K-marginal Subsample Squivalent
- K-marginal PUMA-specific Score
- Univariate Distribution Comparison
- Kendall Tau Correlation Differences
- Pearson Pairwise Correlation Differences
- Linear Regression (EDU vs PINCP_DECILE), with Full 16 RACE + SEX Subpouplation Breakdowns
- Propensity Distribution
- Pairwise Principle Component Analysis (Top 5)
- Pairwise PCA (Top 2, with MSP = 'N' highlighting)
- Inconsistencies (Age-based, Work-based, Housing-based)
- Worst Performing PUMA Breakdown (Univariates and Correlations)
- Privacy Evaluation: Unique Exact Match Metric
- Privacy Evaluation: Apparent Match Metric

E.1 Deidentified Data Summary Table

For convenience, we include the deidentified data summary table from the main paper.

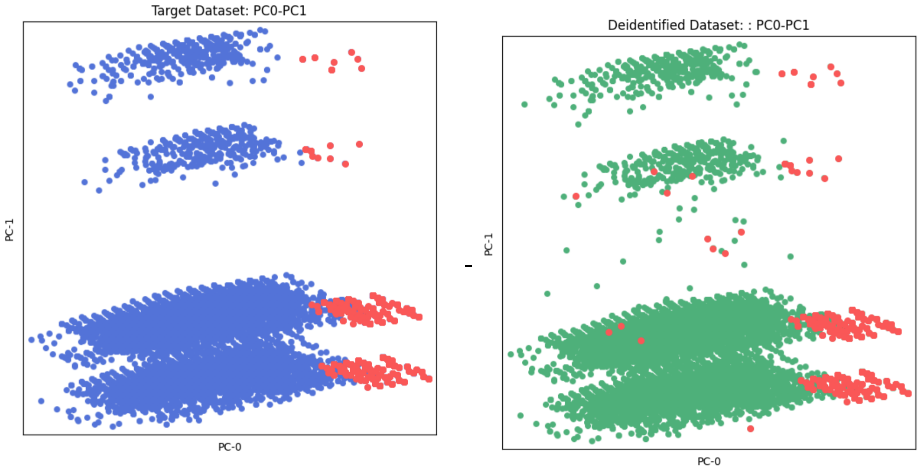
Library and Algorithm	**Privacy Type**	**Algorithm Type**	**Priv. Leak (UEM)**	**Utility (ES)**
DP Histogram (ϵ10)	Differential Privacy (DP)	simple histogram	100%	$\sim$ 90% (988)
R synthpop CART model [2]	non-DP synthetic data	multiple imputation decision tree	2.54%	$\sim$40% (935)
MOSTLY AI SDG [3] [4]	non-DP synthetic data	proprietary pre-trained neural network	0.03%	$\sim$30% (921)
SmartNoise MST (ϵ10) [5]	DP	probabilistic graphical model (PGM)	13.6%	= 10% (969)
SDV CTGAN [6] [7]	non-DP synthetic data	generative adversarial network (GAN)	0.0%	$\sim$5% (775)
SmartNoise PACSynth (ϵ10) [8]	DP + k-anonymity	constraint satisfaction	0.87%	$\sim$1% (551)
synthcity ADSGAN [9] [10]	custom noise injection	GAN	0.0%	< 1% (121)

Table 1: Summary of selected deidentification algorithms. Unique Exact Match (UEM) is a simple privacy metric that counts the percentage of singleton records in the target that are also present in the deidentified data; these uniquely identifiable individuals leaked through the deidentification process. The Equivalent Subsample (ES) utility metric uses an analogy between deidentification error and sampling error to communicate utility; a score of 5% indicates the edit distance between the target and deidentified data distributions is similar to the sampling error induced by randomly discarding 95% of the data. Edit distance is based on the k-marginal metric for sparse distributions. [11], [12]

E.2 Differentially Private Histogram (epsilon-10)

A differentially private histogram is a naive solution that simply counts the number of occurrences of each possible record value, and adds noise to the counts. We use the Tumult Analytics library to efficiently produce a DPHistogram with a very large set of bins. Epsilon 10 is a very weak privacy guarantee, and this simple algorithm provides very poor privacy in these conditions. The points in the 'deidentified' PCA are nearly the exact same points as in the target PCA.

The full metric report can be found here.



(a) Target data (Demographic-focused Feature Subset, excepting AGE and DEYE)

(b) Deidentified data

Figure 2: The PCA Metric for DP Histogram ($\epsilon = 10$)

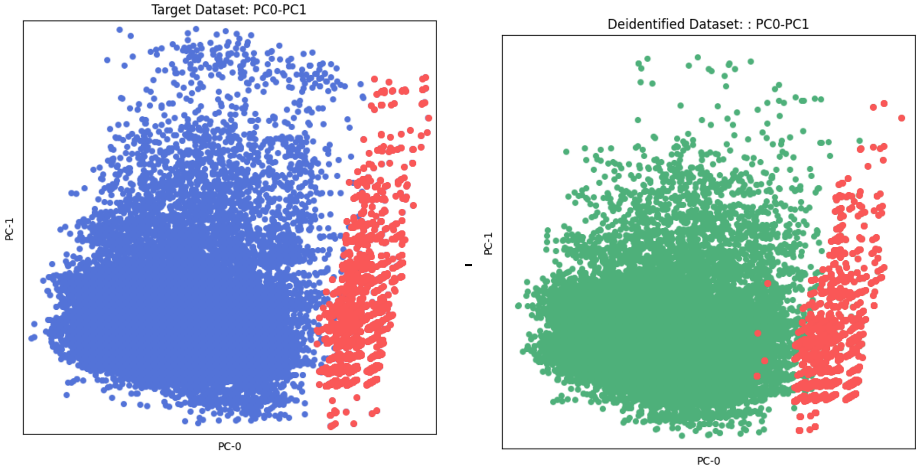
Label Name	Label Value
Algorithm Name	DPHist
Library	Tumult Analytics
Feature Set	demographic-focused-except-AGEP-DEYE
Target Dataset	national2019
Epsilon	10
Privacy	Differential Privacy
Filename	dphist_e_10_cf8_na2019
Records	27314
Features	8
Library Link	https://docs.tmlt.dev/analytics/latest/

Table 2: Label Information for Differential Private Histogram (epsilon-10)

E.3 SmartNoise PACSynth (epsilon-10, Industry-focused)

We've included two samples from the PACSynth library to showcase its behavior on different feature subsets. The technique provides both differential privacy and a form of k-anonymity (removing rare outlier records). This provides very good privacy, Table 1, but it can also erase dispersed subpopulations. The industry feature subset below was used for the regression metric in the main paper, which showed erasure of graduate degree holders among both white men and black women.

More information on the technique can be found here. The full metric report can be found here.



(a) Target data (Industry-focused Feature Subset)

(b) Deidentified data

Figure 3: The PCA Metric for PACSynth ($\epsilon = 10$)

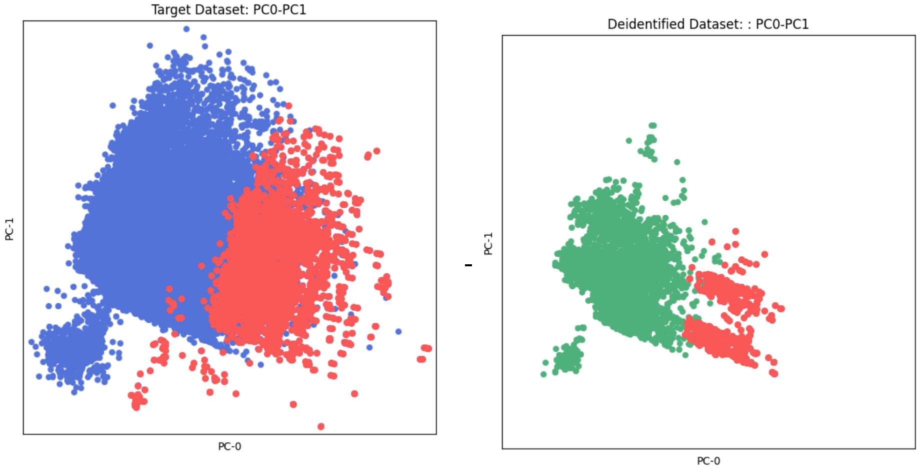
Label Name	Label Value
Algorithm Name	pacsynth
Library	smartnoise-synth
Feature Set	industry-focused
Target Dataset	national2019
Epsilon	10
Variant Label	preprocessor-epsilon: 3
Privacy	Differential Privacy
Filename	pac_synth_e_10_industry_focused_na2019
Records	29537
Features	9
Library Link	https://github.com/opendp/smartnoise-sdk/tree/main/synth

Table 3: SmartNoise PACSynth (epsilon-10)

E.4 SmartNoise PACSynth (epsilon-10), Family-focused)

On the family-focused feature subset we can see the impact of the k-anonymity protection more dramatically. Because the deidentifed data with removed outliers has reduced diversity, it occupies a much smaller area in the plot as compared to the target data. The deidentified records are concentrated into fewer, more popular feature combinations and thus their points show less variance along the PCA axes.

More information on the technique can be found here. The full metric report can be found here.



(a) Target data (Family-focused Feature Subset)

(b) Deidentified data

Figure 4: The PCA Metric for PACSynth ($\epsilon = 10$)

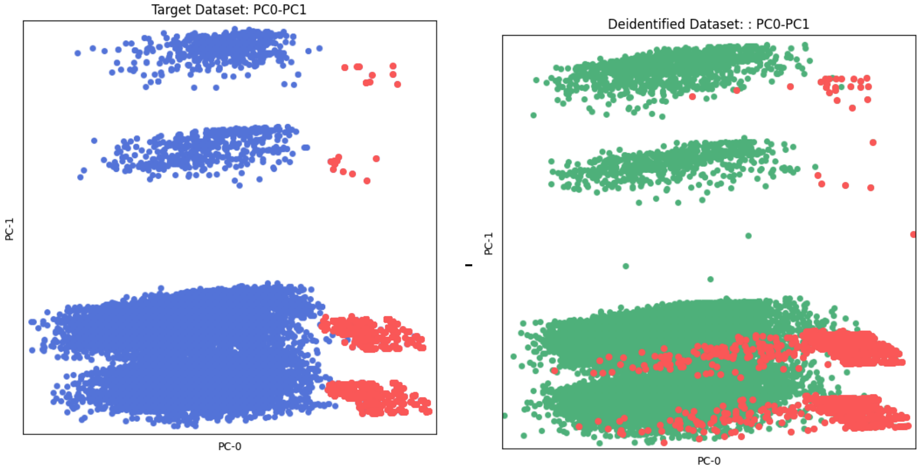
Label Name	Label Value
Algorithm Name	pacsynth
Library	smartnoise-synth
Feature Set	family-focused
Target Dataset	national2019
Epsilon	10
Variant Label	preprocessor-epsilon: 3
Privacy	Differential Privacy
Filename	pac_synth_e_10_industry_focused_na2019
Records	29537
Features	9
Library Link	https://github.com/opendp/smartnoise-sdk/tree/main/synth

Table 4: SmartNoise PACSynth (epsilon-10)

E.5 SmartNoise MST (epsilon-10)

The MST synthesizer uses a probabilistic graphical model (PGM), with a maximum spanning tree (MST) structure capturing the most significant pair-wise feature correlations in the ground truth data as noisy marginal counts. This solution was the winner of the 2019 NIST Differential Privacy Synthetic Data Challenge. Note that it provides good utility with much better privacy than the simple DP Histogram, but its selected marginals fail to capture some constraints on child records (in red).

More information on the technique can be found here. The full metric report can be found here.



(a) Target data (Demographic-focused Feature Subset)

(b) Deidentified data

Figure 5: The PCA Metric for MST ($\epsilon = 10$)

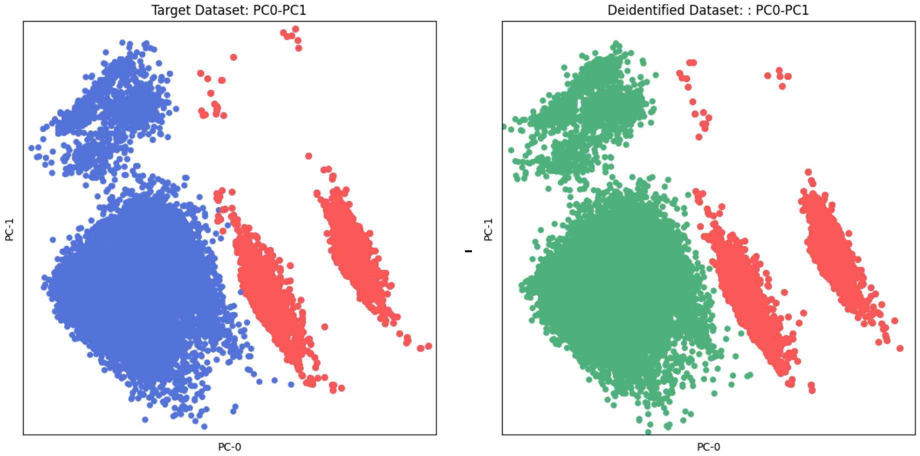
Label Name	Label Value
Algorithm Name	mst
Library	smartnoise-synth
Feature Set	demographic-focused
Target Dataset	national2019
Epsilon	10
Variant Label	preprocessor-epsilon: 3
Privacy	Differential Privacy
Filename	mst_e10_demographic_focused_na2019
Records	27253
Features	10
Library Link	https://github.com/opendp/smartnoise-sdk/tree/main/synth

Table 5: Label Information for SmartNoise MST (epsilon-10)

E.6 R synthpop CART model

The fully conditional Classification and Regression Tree (CART) model does not satisfy formal differential privacy, but provides better privacy than some techniques which do (Table 1). It uses a sequence of decision trees trained on the target data to predict each feature value based on the previously synthesized features; familiarity with decision trees is helpful for tuning this model. Note that the two PCA distributions have very similar shapes, comprised of different points.

You can find more information on the technique here. The full metric report can be found here.



(a) Target data (21-Feature Subset)

(b) Deidentified data

Figure 6: The PCA Metric for CART

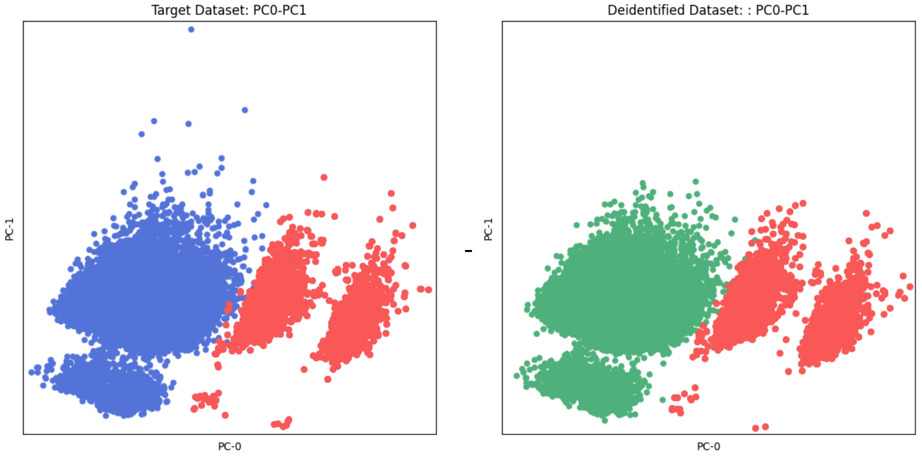
Label Name	Label Value
Algorithm Name	cart
Library	rsynthpop
Feature Set	custom-features-21
Target Dataset	national2019
Variant Label	maxfaclevels: 300
Privacy	Synthetic Data (Non-differentially Private)
Filename	cart_cf21_na2019
Records	27253
Features	21
Library Link	https://cran.r-project.org/web/packages/synthpop/index.html

Table 6: Label Information for R synthpop CART model

E.7 MOSTLY AI Synthetic Data Platform

MOSTLYAI is a proprietary synthetic data generation platform which uses a partly pretrained neural network model to generate data. The model can be configured to respect deterministic constraints between features (for a comparison, see MOSTLYAI submissions 1 in the CRC Data and Metrics Bundle linked above). It does not provide differential privacy, but does very well on both privacy and utility metrics (Table 1).

More information on the technique can be found here. The full metric report can be found here.



(a) Target data (All Features)

(b) Deidentified data

Figure 7: The PCA Metric for MostlyAI

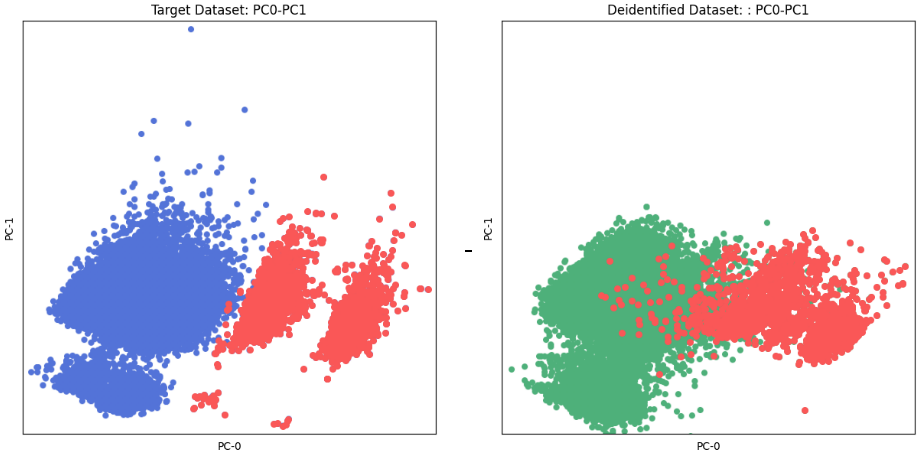
Label Name	Label Value
Algorithm Name	MOSTLY AI
Submission Number	2
Library	MostlyAI SD
Feature Set	all-features
Target Dataset	national2019
Variant Label	national2019
Privacy	Synthetic Data (Non-differentially Private)
Filename	mostlyai_sd_platform_MichaelPlatzer_2
Records	27253
Features	24
Library Link	https://mostly.ai/synthetic-data

Table 7: Label Information for MOSTLY AI Synthetic Data Platform

E.8 Synthetic Data Vault CTGAN

CTGAN is a type of Generative Adverserial Network designed to operate well on tabular data. Unlike the MostlyAI neural network (which is pretrained with public data), the CTGAN network is only trained on the target data. It is able to preserve some structure of the target data distribution, but it introduces artifacts. In other metrics, we see it also has difficulty preserving diverse subpopulations.

More information on the technique can be found here. The full metric report can be found here.



(a) Target data (All Features)

(b) Deidentified data

Figure 8: The PCA Metric for CTGAN

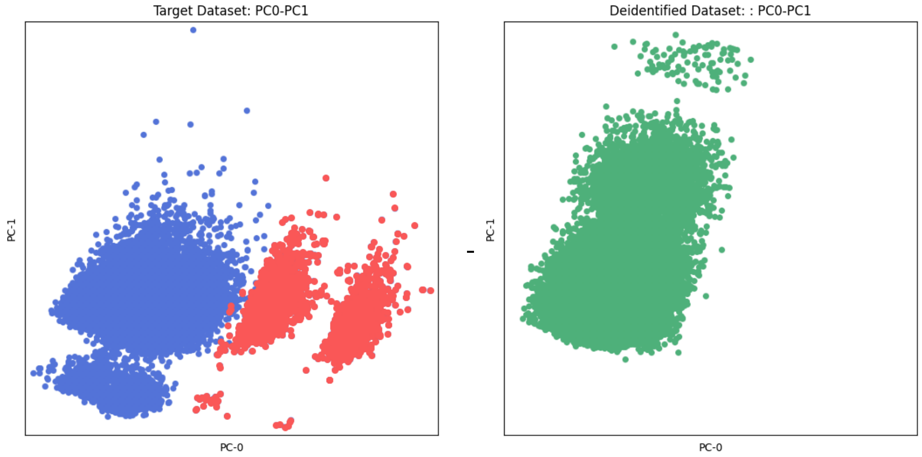
Label Name	Label Value
Team	CBS-NL
Algorithm Name	ctgan
Submission Timestamp	4/16/2023 12:03:58
Submission Number	1
Library	sdv
Feature Set	all-features
Target Dataset	national2019
Variant Label	default CTGAN with epochs=500
Privacy	Synthetic Data (Non-differentially Private)
Filename	sdv_ctgan_epochs500_SlokomManel_1
Records	27253
Features	24
Library Link	https://github.com/sdv-dev/CTGAN

Table 8: Label Information for Synthetic Data Vault CTGAN

E.9 synthcity ADSGAN

ADSGAN is a Generative Adverserial Network focused on providing strong privacy for synthetic data. While it doesn't formally satisfy differential privacy it uses a parameter alpha to inject noise during the training process. Unfortunately, we see it is unable to preserve any meaningful structure from the target data distribution in this submission.

More information on the technique can be found here. The full metric report can be found here.



(a) Target data (All Features)

(b) Deidentified data

Figure 9: The PCA Metric for ADSGAN

Label Name	Label Value
Team	CCAIM
Submission Timestamp	3/9/2023 3:33:23
Submission Number	1
Algorithm Name	adsgan
Library	synthcity
Feature Set	all-features
Target Dataset	national2019
Variant Label	default, lambda=10
Privacy	Synthetic Data (Non-differentially Private)
Filename	adsgan_ZhaozhiQian_1
Records	21802
Features	24
Library Link	https://github.com/vanderschaarlab/synthcity

Table 9: Label Information for synthcity ADSGAN

References

[1] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[2] Beata Nowok, Gillian M. Raab, and Chris Dibben. synthpop: Bespoke creation of synthetic data in r. *Journal of Statistical Software*, 74(11):1–26, 2016. URL: https://www.jstatsoft.org/index.php/jss/article/view/v074i11.

[3] Mostly.ai. Mostly ai synthetic data platform. https://mostly.ai/synthetic-data-platform, 2023.

[4] Michael Platzer and Ivona Krchova. Rule-adhering synthetic data – the lingua franca of learning, 2022. URL: https://arxiv.org/abs/2209.06679.

[5] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data, August 2021. arXiv:2108.04978 [cs]. URL: http://arxiv.org/abs/2108.04978.

[6] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The Synthetic Data Vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, 2016. URL: https://ieeexplore.ieee.org/document/7796926.

[7] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling Tabular data using Conditional GAN. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf.

[8] OpenDP. Private aggregate seeded from pac-synth. https://docs.smartnoise.org/synth/synthesizers/pac_synth.html, 2023.

[9] Jinsung Yoon, Lydia N. Drumright, and Mihaela Van Der Schaar. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, August 2020. URL: https://ieeexplore.ieee.org/document/9034117/.

[10] Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities, 2023. URL: https://arxiv.org/abs/2301.07573.

[11] John Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Daniel Kifer, Philip Leclerc, William Sexton, Ashley Simpson, Christine Task, and Pavel Zhuravlev. An Uncertainty Principle is a Price of Privacy-Preserving Microdata, October 2021. URL: http://arxiv.org/abs/2110.13239.

[12] Grégorie Lothe, Christine Task, Isaac Slavitt, Nicolas Grislain, Karan Bhagat, and Gary S Howarth. SDNist v1.3: Temporal Map Challenge Environment, December 2021. URL: https://data.nist.gov/od/id/mds2-2515.