Federal Register Notice 87 FR 35250, https://www.federalregister.gov/documents/2022/06/09/2022-12432/request-for-information-on-advancing-privacy-enhancing-technologies, June 9, 2022

Request for Information (RFI) on

Advancing Privacy Enhancing Technologies

Knexus Research Corporation

Inequity Resulting from Partitioned Deidentification

Subject: RFI Response: Privacy-Enhancing Technologies

Agency: Office of Science and Technology Policy

RFI Description: The Office of Science and Technology Policy (OSTP) requests public comments to help inform development of a national strategy on privacy-preserving data sharing and analytics, along with associated policy initiatives.

Submission Date: July 8, 2022

Submitted To: PETS-RFI@nitrd.gov

Respondent Organization: Knexus Research Corp.
174 Waterfront Street, Suite 310
National Harbor, MD 20745

Respondent Type: Industry: Gov 8(a) Small business contractor

Respondent Description: KRC is an 8(a)-certified small business and a leading provider of Artificial Intelligence (AI), Privacy-Enhancing Technology (PET) and professional services to the US Government and the commercial sector. KRC has been a performer on Federal Contracts for 16 years since its inception in 2006, and its profile can be searched on the System of Award Management (SAM).

Respondent Technical POC Dr. Christine Task

Respondent Administrative POC Dr. Kalyan Gupta, President

Response Co-Authors: Dr. Christine Task, Lead Privacy Scientist
Aniruddha Sen, Intern

Topics Addressed:
1. Specific research opportunities to advance PETs
2. Specific technical aspects or limitations of PETs
8. Existing best practices that are helpful for PETs adoption

KNEXUS RESEARCH CORPORATION

Inequity Resulting from Partitioned Deidentification

Introduction

Knexus Research Corporation (KRC) has worked with the National Institute of Standards and Technology (NIST) on benchmarking and evaluation of synthetic data generators, the UNECE Synthetic Data Working Group on defining best practices for synthetic data, and the US Census Bureau (USCB) on the research, development, and engineering of synthetic data generators. For this RFI, we wanted to share resources relevant to data deidentification and contribute a simple but important observation about the deidentification of diverse populations.

We begin by defining data deidentification and synthetic data as a category of PET and providing pointers to resources for both understanding the current state of practice and supporting future R&D in this field. [Topic 1]

We then identify a particular inequity issue that can arise due to distributional differences when deidentifying diverse populations. We provide a toy illustration showing that even a relatively large subpopulation (25%) can potentially be unintentionally erased from the data during deidentification, if its distribution differs too significantly from that of the majority population. This applies to all deidentification techniques that use partitioning as a preliminary step, including both traditional deidentification techniques (subsampling) and more recent formal privacy methods (differentially private histograms). [Topic 2]

We recommend, as best practice, that techniques should be evaluated separately on subpopulations as well as the overall population; these evaluations should be considered by stakeholders when configuring a data deidentification approach. [Topic 8]

Finally, we note complexities in this recommendation—distributionally distinct subpopulations may not be easily defined, similar inequity issues may arise more subtly in more complex (non-partitioning) deidentification approaches, and the problem of optimally addressing these issues for diverse populations is not fully solved. We recommend further research.

Deidentified and Synthetic Data

Topic 1: Specific research opportunities to advance PETs

Deidentified Data refers to anonymized data that has been processed with the intention of preventing the reidentification of the individuals in the dataset. Effective deidentification approaches provide both good *privacy* (successful defense against individual reidentification) and good *utility* (query results similar to the original data for population-level queries). Deidentification supports the following capabilities:

- Safe public release of sensitive data, enabling broad transparency and access.
- Safe internal use of sensitive data, reducing the risk of misuse and the burden on cybersecurity to prevent leaks.

Synthetic Data is a category of data deidentification that goes further by using modeling techniques to replace the original population with a new set of artificially generated synthetic individuals who have a similar data distribution at the population level. Because this new synthetic data does not contain real persons, we can refer to it as depersonalized data. It provides additional capabilities:

- Strong individual privacy protection.
- Retention of a data product for population-level analysis after the deletion deadline of the original personal data.
- Reuse of a data product for population-level analysis beyond the application scope of the original personal data.

We believe deidentified data plays an important role in the data privacy ecosystem. However, this is an evolving technology, and research and engineering challenges remain with regards to both privacy and utility. We provide references to the following resources on the current state of deidentification and recent research:

- **UNECE HLG-MOS Synthetic Data Test-Drive Website**
  - Over twenty national statistical agencies and interested organizations around the globe participated in a test drive of currently available synthetic data techniques.
  - Resource type: Directory of currently used synthesis techniques, as well as utility and privacy metrics chosen by data stakeholders. Challenge submissions include discussion of successes and failures for current techniques.
  - Location: https://pages.nist.gov/HLG-MOS_Synthetic_Data_Test_Drive/
- **NIST Differential Privacy Synthetic Data Challenges**
  - A sequence of eight challenge sprints over four years, scoring the performance of formally private synthetic data generators in difficult, real-world data contexts.
  - Resource type: Archive of challenge problems, techniques, performance results, and open-sourced solutions.
  - Location: https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2020-differential-privacy-temporal

Partitioning Plus Privacy Impacts Equity

Topic 2: Specific technical aspects or limitations of PETs

Data deidentification is intended to protect individual privacy and support population-level data analysis. As a result, it doesn't work on small groups--if a subgroup contains only a few individuals, such that they are recognizable as clearly distinct in the data, then as part of ensuring that no individual can be reidentified, the deidentification process will generally ensure poor utility for queries that would single out these individuals.

This raises an important question: What makes a group small?

We briefly introduce one behavior of deidentification algorithms on diverse populations. A diverse population is defined here to be a dataset containing more than one subpopulation, in which each subpopulation has a significantly differing distribution in the feature space. We draw a distinction (not typically done) between deidentification techniques that partition the individuals in the dataset and those that do not:

- **Partitioning Deidentification (table-based):** These approaches begin by considering the data in table form, where each column represents a feature (e.g., sex, race, or county) and each row has the count of individuals with a given combination of feature values (e.g., [Female, White, Alexandria]: 25,312). Deidentification is then performed by altering these counts. Because each individual contributes to only one count, we say that these techniques *partition* the data. Partitioning deidentification includes traditional approaches, such as

subsampling and cell suppression, as well as some recent techniques, such as differentially private histograms.

- **Non-partitioning Deidentification (query-based):** Non-partitioning approaches capture the data distribution using a sequence of queries in which one individual may contribute to multiple query results; different queries typically focus on different aspects of the data distribution. These techniques include model-based synthetic data.

An implicit assumption of partitioning-based deidentification is that small table counts represent outliers in the data, who (1) are more vulnerable to reidentification due to their unusual combination of feature values, and (2) as outliers can be altered or removed without substantially impacting analyses of interest. Deidentification then protects these individuals by using approaches such as redacting them (e.g., k-anonymity, cell suppression) or randomizing their counts (e.g., subsampling, differentially private histograms).

However, this assumption may be overly simplistic. In diverse populations, the choice of which features to include in the schema can potentially cause a relatively large subpopulation to be dispersed across many small table counts. This could result in the subpopulation being heavily altered or erased during deidentification. We illustrate this with a toy example below.

Figure 1: Subsampling

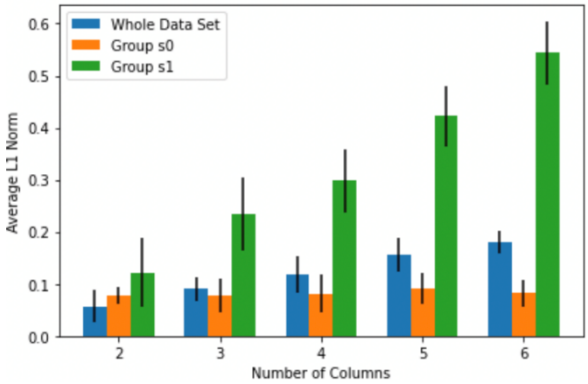
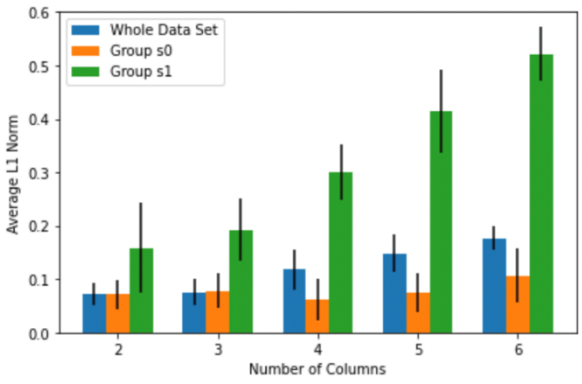


Figure 2: Simple DP Histogram



Our toy example contains a majority subpopulation s0 comprising 75% of the data, and a minority subpopulation s1 comprising the remaining 25%. We consider two partition-based deidentification algorithms:

- **Subsampling and weighting:** A 50% subsample of the original data is released, with each record given a weight of 2. Subsampling is one component of the privacy solution currently in use for the American Community Survey microdata
- **DP Histogram:** Laplacian noise with scale 1 (epsilon = 1) is added to every table count. This is arguably the simplest approach for differentially private data sharing and provides formal privacy guarantees against the release of individually identifying information.

Our experiment begins with a schema containing only two features (e.g., age and sex) and then iteratively adds additional features by appending new columns to the table. Each new column adds new information about the individuals in the data set. In general, releasing more information means individuals are more identifiable, and this increases the impact of deidentification on utility.

However, because the two subpopulations have different distributions, their utility is impacted differently. We use an extreme case for this example: In subpopulation s0 the new columns are strongly correlated with previous columns, whereas for subpopulation s1, the new columns are independent of the previous columns. This means that the distribution of the majority subpopulation will tend to remain concentrated and preserved with the addition of the new columns, whereas the

distribution of the minority will tend to be diluted and dispersed. When all six columns are added, the minority subpopulation becomes spread across many small table counts; the new features are effectively more identifying for s1 than s0.

Figures 1 and 2 show the average error (expressed as L1 norm between the original and deidentified data density) while the number of columns increases. Error is shown separately for each subpopulation and for the group as a whole. In this toy example, we see that the error of the minority s1 increases dramatically, whereas the error of the majority s0 remains essentially constant. The overall population error only gradually increases as new information is added.

This type of utility inequity has the following properties:

- It arises when subpopulations have significantly differing distributions. It is especially a concern for partitioning deidentification algorithms, including traditional deidentification techniques.
- There are many ways this can occur beyond the example above; in general, a given partitioning scheme can have a disparate impact when subdividing different subpopulations.
- It can potentially impact even relatively large subpopulations (25% of the population in the above example).
- It may not show up clearly in utility evaluations that cover the full population, and so it may be overlooked when the deidentification approach is being designed/configured. This is true both when it's a person deciding what information to release, or an algorithm fitting a model (more below).
- It should be addressed by separately considering utility for each subpopulation.

Moving beyond our toy example, we'd like to add three more points:

- We expect this issue may also arise to varying extents in more complex data-deidentification approaches, including non-partitioning approaches. Query results, dimension reduction techniques, summary statistics and model training are still dependent on choice of schema.
- Importantly, subpopulations in the data may not be trivial to identify a priori. It is always valuable to evaluate with respect to basic demographic divisions, but distributionally differing subpopulations may also be defined by cultural, work, community, or lifestyle attributes that aren't captured by race or gender.
- We believe this topic requires further research.

Subpopulation Evaluation as Best Practice for Maintaining Equity

Topic 8: Existing best practices that are helpful for PETs adoption

For data deidentification, we recommend that utility for subpopulations be evaluated separately, in addition to overall utility, and that privacy and fairness be considered together when stakeholders are deciding on a data release strategy. We recommend this for all data deidentification approaches, including recent techniques (differential privacy, synthetic data) and techniques in common use currently (subsampling and cell suppression).

The previously referenced HLG-MOS Synthetic Data Test-drive website contains a large directory of utility evaluations for deidentified data. Additionally, in the coming months the SDNist library is expecting to release a set of benchmark data and evaluation tools designed to support detailed exploration of algorithm performance on diverse, real world populations. Updates will be posted here: https://github.com/usnistgov/SDNist