Formally Private Synthetic Data:
Empowering research on confidential data while providing for privacy and security

Simson L. Garfinkel
Center for Disclosure Avoidance
US Census Bureau[[1]](#footnote-1)

March 16, 2017

Many organizations have confidential data that can be prohibitively complicated or expensive to share with external researchers for reasons of law, security, or logistics. *Formally private synthetic data* *(FPSD)* is an approach that these organizations can use to enable external researchers to perform a wide variety of aggregate analyses on the confidential datasets while maintaining limits on the amount of privacy loss that individuals within the dataset can experience. Compared with other techniques for allowing statistical analyses of confidential data, such as de-identification and secure computing enclaves, formally private synthetic data potentially offers stronger privacy and security guarantees, but at a cost of data accuracy and public acceptability. Additionally, FPSD cannot be linked with external datasets, so they cannot be used for linkage studies or certain kinds of longitudinal studies.

Synthetic data have long been used in social science research, as they allow students and researchers to become familiar with a dataset without having to obtain approvals necessary to work with confidential data. Synthetic data have also been used in software development, as it allows programmers to create and test software without the additional security concerns that come with handling personally identifiable information (PII).

Synthetic data can be created by two approaches:

* Sampling an existing dataset and either adding noise to specific cells likely to have a high risk of disclosure, or replacing these cells with imputed values. (A “partially synthetic dataset.”) These kinds of datasets do not offer formal privacy guarantees.
* Using an existing dataset to create a model that is, in turn, used to create a synthetic dataset through some kind of stochastic process. (A “fully synthetic dataset.”) These kinds of datasets can offer formal privacy guarantees, depending on how the model is created.

In both cases, formal privacy techniques such as Differential Privacy can be used to quantify the privacy protection offered by the synthetic dataset.

Model Generation

Confidential
Data

Model with Privacy Guarantees

Synthetic Data Generation

Formally Private Synthetic Data

FPSD can help statistical agencies realize their legal requirements of not publicly releasing information that could identify a respondent while still allowing researchers to perform a variety of statistical analyses on the dataset:

* It can be very difficult or even impossible to map the synthetic data back to actual people. Thus, unlike Public Use Microdata Samples (PUMS), FPSD has substantially reduced and fully controlled risk of record re-identification, because it can only leak information about individuals to the extent allowed under the formal privacy guarantees.
* The privacy guarantees can potentially be mathematically established and proven (cf. the section below on “Creating a synthetic dataset with differential privacy”).
* The privacy guarantees can remain in force even if there are future data releases, authorized or unauthorized.

FPSD is an exciting technique that is presently making its way from the research lab and into practical applications. For example, a hospital could use a FPSD approach to create a dataset of hospital admissions. Such a dataset, based on actual patient admissions and outcomes, could be made available for download on the hospital website without the need for vetting researchers: since the dataset did not contain any actual patient information, it would likely not be considered protected health information (PHI) under the Safe Harbor provisions of the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, because the dataset would not contain any of the 18 proscribed identifiers for the hospital’s patients. Researchers, students, and ordinary members of the public could download the dataset and learn about healthcare utilization costs, build statistical models predicting outcomes from treatments, and even do original research about hospital acquired infections. But a journalist who downloaded the data and tried to discover the actual patients who had acquired clostridium difficile (C. diff) in Room 305 would be out of luck, those specific records would not be from actual real people—they would have been synthesized from a formally private model. Likewise, a medical researcher attempting to link those C. diff infections to a specific piece of medical equipment, or a work shift of a particular nursing assistant, would similarly would find the publicly released data insufficient for that purpose.

FPSD can also be the first part of an integrated approach for allowing researchers to work with highly sensitive confidential data:

* Researchers can perform exploratory work and develop analytics approaches using the synthetic datasets.
* Once an analysis program is developed, the researchers could provide the analysis program to a *verification or validation server* at the host agency. The verification server runs the program on both the original data and the synthetic data. If the results of the two runs are substantially similar, the researcher can be told this and they are free to publish the results based on the synthetic data. But if the results are substantially different, the results based on the real data could be reviewed by a disclosure review board (DRB), which would determine if the results of the analysis on the real data could be published or if they would need to undergo some other form of statistical disclosure limitation.
* Finally, if the researcher wishes to perform exploratory research, they could submit an application based on their preliminary findings with synthetic data to be granted access to the real data. They could then travel to a *secure enclave* and work with the original raw data. Once the researcher has results, those results would need to be submitted to the DRB for release approval.

As this example shows, FPSD can have an important role in evidenced based research. But for that to happen, more work needs to be done:

1. Although synthetic data has been used for years within the social sciences, there is little experience and few off-the-shelf algorithms for creating synthetic data that is protected for formal privacy guarantees.
2. It is difficult to find meaningful correlations or abnormalities in the synthetic data that are not represented in the model. For example, if a model contains only main effects and first-order interactions, then all second-order interactions can only be estimated from the synthetic data to the extent that their design is correlated with the main or first-order interactions. Thus, model creation is of primary importance when creating a synthetic data set, and problems in the model (and the resulting synthetic data) may not be apparent at the time of creation.
3. Users of the data may not realize that the data are synthetic. Simply providing documentation that the data are fully synthetic may not be sufficient public notification, since the dataset may be separated from the documentation. Instead, the datasets should be labeled internally.
4. Releasing a “synthetic” dataset may not be regarded by the public as a legitimate tool for promoting accountability or transparency. For example, a journalist equipped with a synthetic data release would be unable to find the actual “people” who make up the release, because they would not actually exist. These concerns can be addressed with public education, to explain to potential users the specific uses for synthetic data, and through the development of metrics that convey the accuracy of synthetic data for performing statistical analysis.

To date, various forms of both privacy-protecting synthetic data and formally private synthetic data are in use, including:

* The US Census Bureau’s Survey of Income and Program Participation (SIPP) Synthetic Beta (SSB). This dataset “integrates person-level micro-data from a household survey (SIPP survey data) with W-2 earnings and OASDI benefits data. Unlike the original, administrative data, the SSB is publicly available… The Census Bureau offers all SSB users the opportunity to submit their programs for validation on the internal, confidential version of these data.”[[2]](#footnote-2)
* The US Census Bureau’s Synthetic Longitudinal Business Database (SynLBD) Beta is an experimental synthetic product that provides 21 million synthetic establishment records covering the years 1976-2000. The dataset contains no geographic or firm-level information. The Census Bureau offers a validation service in which SAS or STATA code written for the SynLBD can be run on the actual dataset. The analysis is reviewed by Census Bureau Disclosure Review Officers prior to being made available to external researchers.[[3]](#footnote-3)

References

Privacy-Preserving Data Analysis for the Federal Statistical Agencies, John Abowd, Lorenzo Alvisi, Cynthia Dwork, Sampath Kannan, Ashwin Machanavajjhala, and Jerome Reiter, January 2017, Computing Community Consortium. <https://arxiv.org/pdf/1701.00752.pdf>

Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. [Privacy: From Theory to Practice on the Map.](http://www.cs.cornell.edu/johannes/papers/2008/icde2008-privacy.pdf) In *Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE 2008)*, Cancun, Mexico, April 2008

Li, H., Xiong, L., Zhang, L., & Jiang, X. (2014). DPSynthesizer: Differentially Private Data Synthesizer for Privacy Preserving Data Sharing. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, *7*(13), 1677–1680. http://doi.org/10.14778/2733004.2733059

Vincent Bindschaedler, Reza Shokri, and Carl A. Gunter. 2017. Plausible deniability for privacy-preserving data synthesis. *Proc. VLDB Endow.* 10, 5 (January 2017), 481-492. DOI: https://doi.org/10.14778/3055540.3055542

1. **DISCLAIMER: The views expressed in this document are those of the author and do not necessarily reflect the policy of the US Census Bureau, the US Department of Commerce, or the United States Government.** [↑](#footnote-ref-1)
2. https://www.census.gov/programs-surveys/sipp/events/2016-SIPP-Workshop-SSB.html [↑](#footnote-ref-2)
3. https://www.census.gov/ces/dataproducts/synlbd/validatingresults.html [↑](#footnote-ref-3)