Technical Perspective: Synthetic Data Needs a Reproducibility Benchmark

Xi He
University of Waterloo
xi.he@uwaterloo.ca

Synthetic data is a vital substitute for real sensitive personal data in supporting social science research and policy studies. Extensive prior research has delved into various models for generating synthetic data, from traditional statistical approaches to cutting-edge deep-learning methods. However, selecting the most suitable one for unforeseen applications poses a significant challenge due to the varying strengths and weaknesses, dependent on factors such as the application domain, data distribution, analytical requirements, and privacy considerations.

Differential privacy (DP) synthesizers have emerged as a prominent approach for generating synthetic data, offering strong mathematical privacy guarantees. These synthesizers first learn a model from real data, inject noise to achieve DP, and then sample the noisy model to generate synthetic datasets. Despite DP offering stronger privacy assurances than its predecessors, such as k-anonymity, DP synthesizers face many utility concerns and criticisms. The concerns are particularly pertinent in real-world applications, such as the U.S. Census's 2020 release, where noise in the data could lead to inaccurate or biased outcomes for critical decisions like allocating block grants.

However, do these concerns apply to non-DP synthesizers, such as the census release before 2020? They likely do, especially if they offer comparable privacy protection, yet there is limited evidence regarding the utility of any synthesizers in practical settings. Common utility proxies for synthetic data evaluation, like descriptive statistics and classification accuracy, offer some procedural representation of analysis tasks but lack validation for real-world applicability. Addressing this gap requires the development of realistic utility benchmarks, a pressing and outstanding problem.

In their article "Epistemic Parity: Reproducibility as an Evaluation Metric for Differential Privacy," the authors propose an evaluation methodology for synthetic data centered on a question: "Can a DP synthesizer produce private tabular data that can be used to derive scientific findings?" This methodology directly addresses practitioners' experience with synthetic data in real-world scenarios, measuring the likelihood that published findings would have changed had the authors used synthetic data, a condition termed epistemic parity. The authors begin by reproducing findings found in peer-reviewed papers using real datasets and then replicate these on their DP synthetic version for comparison.

While this idea may seem straightforward, its execution demands significant effort due to various challenges. Firstly, reproducing conclusions from peer-reviewed papers often encounters low success rates due to factors like unclear computational details and data versioning issues. Second, crafting an effective benchmark to capture real-world analysis complexity and ensure fair comparisons among DP synthesizers requires meticulous navigation and filtering of vast datasets and their studies. Furthermore, each DP synthesizer entails multiple sources of randomness, complicating the task of confidently and fairly reporting evaluation scores. This paper is the first to tackle all these challenges and present a practical taxonomy for reproducing statistical analyses in peer-reviewed publications and a software benchmark package, SynRD, that automates the epistemic parity evaluations for DP synthesizers.

The authors leveraged concepts from reproducibility literature to carefully select datasets from ICPSR, a data repository for social science (consisting of over 100,000 publications spanning 17,312 studies), identify conclusions in the papers, extract relevant findings, and implement corresponding statistical tests. The SynRD benchmark comprises eight datasets, each rigorously reviewed by two researchers with expertise in computing science, statistics, or both, entailing a minimum of thirty hours of work. This effort yielded over a hundred reproducible empirical findings. The selected datasets and findings exhibit diverse properties and characteristics (sample size, number of variables, domain size, outliers, mutual information, skewness, and sparsity) and cover various computational methods such as regression, causal paths, etc. Using this new benchmark on five state-of-the-art DP synthesizers, the authors recover the main results from prior benchmark papers that use simple utility proxies. There are also new and surprising insights, such as the low sensitivity to the privacy budget on reproducibility and the failure of all synthesizers for specific challenging datasets.

This open-sourced and extensible benchmark will continue to grow with new characteristics for evaluating synthetic data, such as its false discovery rates and balance between utility and privacy. SynRD has the possibility to become one milestone benchmark that advances DP synthesizers and other practical synthesizers beyond DP research. If you are intrigued by the construction of SynRD and the new insights into the latest DP synthesizers, or if you aspire to develop cutting-edge DP synthesizers or contribute to synthetic data benchmarks, we highly recommend reading this paper.