Evaluating Differentially Private Synthetic Data Generators in Social Science Settings

Kelly Xu
Rutgers University
New Brunswick, New Jersey

Thomas Chen
University of California, Berkeley
Berkeley, California

Ruobin Gong
Rutgers University
New Brunswick, New Jersey

Abstract—**Differential privacy is essential in modern data analysis and sharing. Unlike traditional privacy protection methods, differential privacy ensures the privacy of individuals in a dataset, preventing malicious agents from identifying or inferring specific individual information. However, it faces scalability challenges with large datasets and requires a delicate balance between privacy and data utility.**

In this study, we evaluate current differentially private synthetic data generators in recent social science studies using large public datasets. We focus on the synthetic data generators DataSynthesizer and PrivBayes, analyzing their effectiveness in replicating results from the Current Population Survey. Our findings indicate that DataSynthesizer struggles with complex queries that consider more than just marginal distributions. Additionally, we examine how the quality of synthetic data varies across different settings, providing future guidance on producing high-quality differentially private synthetic datasets.

I. Introduction

Data privacy has become a critical issue in the modern digital era, driven by the exponential growth of data generation and the increasing ability to store, process, and analyze large datasets. The richer the data in a dataset, the more useful it is for analysis. Historically, people used the practice of removing personally identifiable information (PII), such as names, to anonymize data and protect data privacy. However, the problem with traditional anonymization is that the richness of the data still allows adversaries to identify individuals within the dataset. A classic example of this issue dates back to 1997. As discussed by Daniel Barth-Jones in an article for SSRN Electronic Journal, a graduate student identified the medical records of then-Governor of Massachusetts William Weld within an anonymized dataset. She achieved this by comparing voter registration records, which included zip codes, birth dates, and gender, with the anonymized data, thereby identifying Weld's medical records.

Although there are now stricter controls on the release of privacy-sensitive data like medical records, many sources of raw data remain readily accessible. Compared to traditional anonymization methods, differential privacy ensures that removing a single data point does not significantly alter the overall dataset, thereby protecting user privacy. Differential privacy can help users feel secure when participating in surveys, responding to census questions, or allowing hospitals to share medical information for scientific use. It has become a superior solution, providing a robust means of achieving privacy protection in today's data-rich environment.

Recently, researchers have considered the combination of differential privacy (DP) and data synthesis as a solution for releasing analytically useful data while preserving the privacy of individuals in the data. To utilize these algorithms for public policy decisions, policymakers need an accurate understanding of the comparative performance of these algorithms. However, there are very few studies comparing multiple differentially private data synthesis methods, and none have applied these comparisons to real-world data. Therefore, our goal is to learn how to generate synthetic datasets that effectively preserve usability while satisfying differential privacy in real-world datasets. Our first step was to investigate different types of differential privacy synthetic data generators and consider which generator to implement and test. We referenced Bowen and Snoke's ”Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge,” which analyzed various differential privacy synthetic data generators that participated in the 2019 NIST PSCR Differential Privacy Synthetic Data Challenge. Ultimately, among the numerous differential privacy synthetic data generators, we selected the PrivBayes and DataSynthesizer algorithms because they performed relatively well in the competition and utilized the simplest data preprocessing and postprocessing. Both DataSynthesizer and PrivBayes algorithms are based on Bayesian networks, a probabilistic model that represents the distribution of attributes and their dependencies.

The second step was to consider the dataset and study on which to implement the data generator. We ultimately chose the Current Population Survey (CPS) as our dataset because it provides data from a nationally representative sample of US households of non-institutionalized civilians. CPS is frequently used in labor market analysis, policy evaluation, and social economic research. For the study, we selected a cross-sectional data study that primarily used the CPS-TUS dataset and evaluated the dataset's performance on various queries and statistics. The main study we chose to analyze is ”E-cigarette Use and Associated Changes in Population Smoking Cessation: Evidence from US Current Population Surveys,” published in BMJ in 2017. This study examined the relationship between the increase in e-cigarette use and changes in smoking cessation rates at the population level. The study includes multiple tables that show the proportions of smokers and recent quitters who used e-cigarettes and uses ^{2} tests or normal approximation to ^{2} tests to evaluate the differences in quit attempts and success rates between e-cigarette users and non-users.

II. Definitions

A. Differential Privacy

Differential Privacy(DP) is a mathematical framework to quantify the amount of privacy provided by an algorithm.

Let $\mathcal{X}$ be a public dataset of dimension $n \times q$. Given ϵ, a sanitation algorithm $\mathcal{M}$ is $\epsilon - DP$ for all $S \subset range(\mathcal{M})$ and for all X and X' that differ by one record, it fulfills the following equation:

$$\frac{\Pr(\mathcal{M}(X) \subset S)}{\Pr(\mathcal{M}(X') \subset S)} \leq exp(\epsilon)$$

Many other flavors of DP exist, but this is the primary definition that we will use.

B. Laplace Mechanism

In the data synthetic generators we discuss, we will primarily use the Laplace mechanism to inject noise into the algorithms to satisfy differential privacy. Given a function F, the Laplace Mechanism satisfies DP by adding noise drawn from a Laplace distribution with the location parameter at 0 and scale parameter of $\frac{\delta(F)}{\mu}$, where $\delta(F)$ is the global sensitivity of F.

The global sensitivity of a function F is given by

$$\delta(F) = sup_{X,X'}||F(X) - F(X')||$$

Where $||\cdot||$ is the L_1 norm and X and X' are once again 2 datasets that differ by one record.

C. Bayesian Networks

A Bayesian Network is a directed acyclic graph(DAG) where nodes represent attributes and the edges represent dependencies between attributes. Given a dataset X with d attributes, we can model the dataset as a joint distribution over each attribute domain(Zhang et. al 2017). We can formally model the distribution of records in X with d conditional distributions. Let $X_1, X_2, X_3 \ldots X_d$ be the attributes of the set. Then given a Bayesian network, we have that

$$(X) = \Pi_{i=1}^{d} \Pr[X_i|\Pi_i]$$

where Π_i is the parents of X_i as seen in Zhang et. al 2017.

Therefore, we can use a Bayesian network to model conditional distributions and thus dependencies between attributes when generating synthetic data.

III. PrivBayes and DataSynthesizer

Datasynthesizer and PrivBayes are very similar methods. Both algorithms first generate a Bayesian network based on the attributes with noise injected to satisfy differential privacy. The root of the network is randomly chosen in both algorithms. Then based on the Bayesian network, noisy conditional distributions are created for each attribute. Finally, based on the Bayesian network and noisy conditional distributions, we create a noisy distribution of the original dataset and sample it to get a synthetic dataset. The main difference between the two methods is how they generate the Bayesian Network. Datasynthesizer generates the network based on the maximum possible degree k. Users may choose the maximum degree to set and the algorithm will give every node that many parents if possible. One of the issues with this is that higher-degree networks will inject significantly more noise as increasing the degree increases the domain of the marginal distributions. Since the algorithm must satisfy DP, the privacy budget will be divided over marginal distributions with larger domains, which are highly sensitive to noise. Therefore ϵ needs to be adjusted accordingly for higher degree networks or require a smaller number of attributes so that the noise injected is much lower. PrivBayes uses another variable, θ, to dynamically estimate the maximum number of parents a node can have. θ, defined by Zhang as θ usefulness, essentially measures how useful it is to add a specific parent node. By definition, if the ratio of the average scale of information to the average scale of noise is no less than θ, then it is theta useful(Zhang et al. 2017). This leads to a more varied network compared to the Bayesian network generated by Datasynthesizer. Both methods consider a β variable, which is the percentage of the privacy budget given to generate the Bayesian network. The other $1-\beta$ of the privacy budget will be given to generate the marginal distributions based on the network.

IV. Data and Methods

A. Data

The Current Population Survey (CPS) is a monthly survey conducted by the United States Census Bureau and the Bureau of Labor Statistics. It is a primary source of labor force statistics for the U.S. population and provides comprehensive data on employment, unemployment, earnings, hours of work, and other indicators. We used the data from 2014-15 tobacco use supplement. The sample sizes are 161725. All data are downloaded from the Integrated Public Use Microdata System (IPUMS).

B. Methods

We primarily consider the first table in the E-cigarette study. Table 1 looks at the rates of ever use and current use of e-cigarettes by various demographics based on the 2014-15 US Current Population Survey-Tobacco Use Supplement (CPS-TUS). The table provides the proportions of respondents who have ever used e-cigarettes and those who are current users, separated by characteristics such as sex, age, ethnicity, and education level. Secondly, we checked if the relationship between the increase in e-cigarette use and changes in smoking quit attempts still existed in the datasets after synthesis. The proportions are presented with 95We first replicated the results in the original paper from the 2014-15 CPS-TUS dataset and used the results as our original result. Next, we used DataSynthesizer and PrivBayes to generate synthetic datasets that are the same size as the dataset and then applied the same analysis as the original one to produce table 1 with the synthetic data. In all experiments, the algorithms were repeated 10 times for each setting to observe the variability of the synthetic data. We obtained and used the average values from these repetitions for further analysis. In order to observe the effects of ϵ and k, we first used Datasynthesizer algorithm to generate datasets with different combinations of ϵ and k from the list $\epsilon = 1, 2, 10, 100$ and $k = 1, 2$. To convenience the comparing with PrivBayes algorithm, we set $\beta = 0.3$. We also considered how the PrivBayes algorithm compared to the Datasynthesizer algorithm. We ran the PrivBayes algorithm on the same dataset that Datasynthesizer algorithm used to with $\epsilon = 1, 2, 10, 100$ and $\beta = 0.3$ and $\theta = 4$ (as it recommended and justified by Zhang et. al, 2017), and compared the results to the Datasynthesizer algorithm on the same settings. To evaluate the accuracy of the synthetic data, we compared the mean average of 10 repeated in each setting with the original results. Then we demonstrate both the proportions and their 95

V. Results

A. A.Evaluating the effect of ϵ and k on Table 1

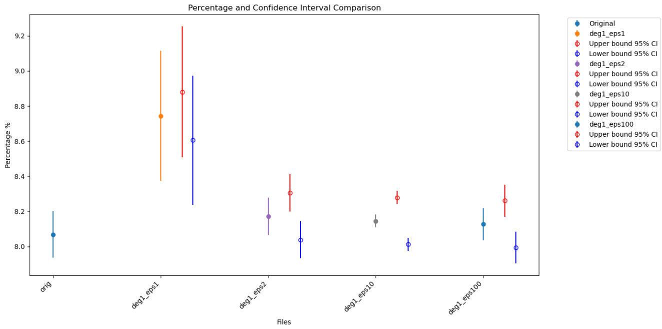

Fig. 1. Percentage of ever use of e-cigarettes with 95% confidence intervals on the original dataset and synthetic datasets generated using DataSynthesizer (Bayesian network with degree $k = 1$ across different privacy budgets ϵ

Here, we chose the overall rates of ever use of e-cigarettes to present because this value is determined by a single variable, ”ECEV,” in the 2014-15 CPS-TUS dataset. This choice allows us to evaluate the effect of ϵ and k and effectively showcases the ability of both PrivBayes and DataSynthesizer to produce correct statistical significance in simple queries.

Fig. 1-2 show the overall rate of ever use of e-cigarettes with 95% confidence intervals on the original and synthetic datasets (generated using DataSynthesizer). The general trend indicates that as ϵ increases, the rates of ever use of e-cigarettes in the synthetic datasets become closer to the original rate.

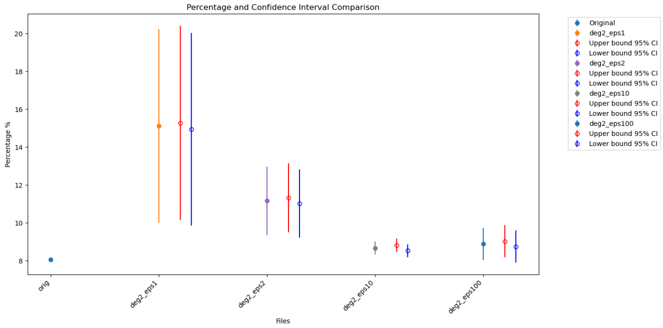

Fig. 2. Percentage of ever use of e-cigarettes with 95% confidence intervals on the original dataset and synthetic datasets generated using DataSynthesizer (Bayesian network with degree $k = 2$ across different privacy budgets ϵ

Comparing the results for $k = 1$ and $k = 2$, we observe that when ϵ is small, the rate on synthetic datasets with $k = 2$ deviates more from the original rate than with $k = 1$. However, as ϵ increases, the rate for $k = 2$ more quickly approaches the original rate. By $\epsilon = 10$, the proportion on the synthetic dataset with $k = 2$ is already very close to the original rate.

B. Comparing the PrivBayes with DataSynthesizer (Table 1)

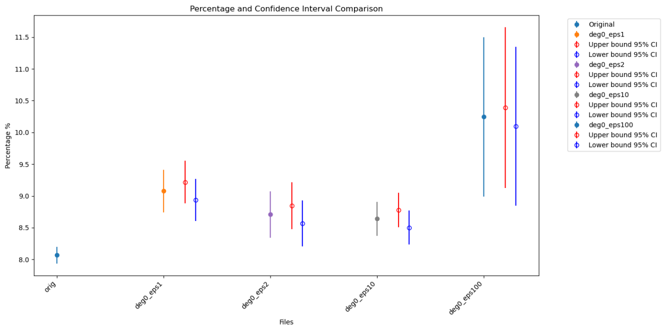

Fig. 3. Percentage of ever use of e-cigarettes with 95% confidence intervals on the original dataset and synthetic datasets generated using PrivBayes under varying privacy budgets ϵ

Fig. 3 shows the overall proportions of ever use of e-cigarettes with 95% confidence intervals on the original and synthetic datasets (generated using PrivBayes). We observed that when ϵ is small, the overall trend is similar to that of DataSynthesizer. However, when $\epsilon = 100$, the proportion on the synthetic datasets shows a significant difference from the original proportion.

C. Evaluating the effect of ϵ and k on the relationship between the increase in e-cigarette use and changes in smoking quit attempts

The quit attempt rates of former e-cigarette users and current users are determined by multiple variables in the 2014-15 CPS-TUS dataset. This choice allows us to evaluate the ability of both PrivBayes and DataSynthesizer to produce correct statistical significance in complex queries.

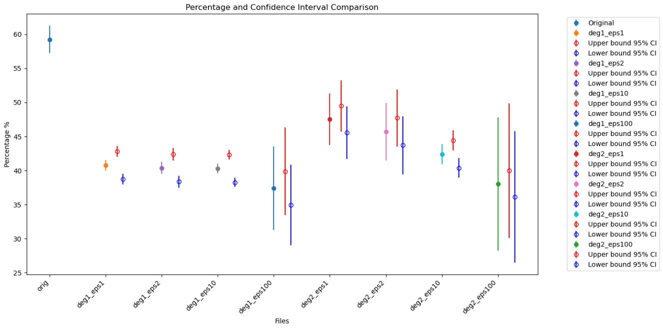

Fig. 4. Smoking quit attempt rates among former e-cigarette users, with 95% confidence intervals shown for both the original dataset and synthetic datasets generated using DataSynthesizer

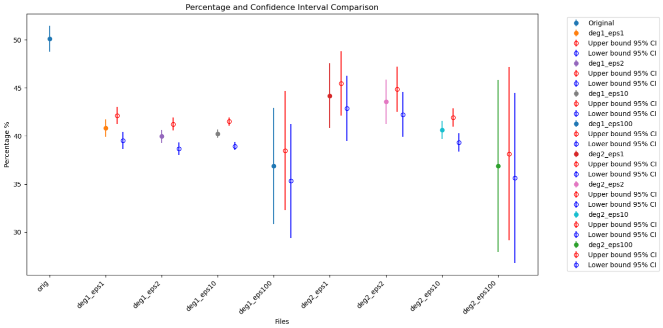

Fig. 5. Smoking quit attempt rates among current e-cigarette users, with 95% confidence intervals shown for both the original dataset and synthetic datasets generated using DataSynthesizer

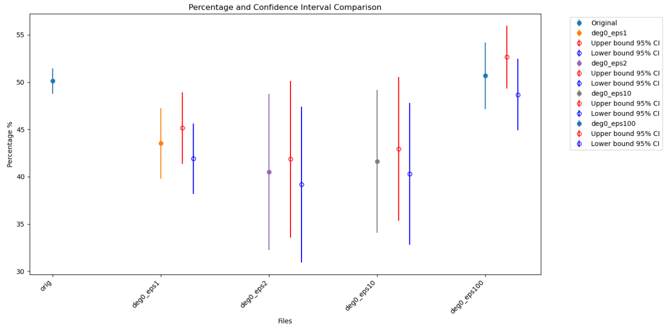

Fig. 6. Smoking quit attempt rates among former e-cigarette users, with 95% confidence intervals shown for both the original dataset and synthetic datasets generated using PrivBayes

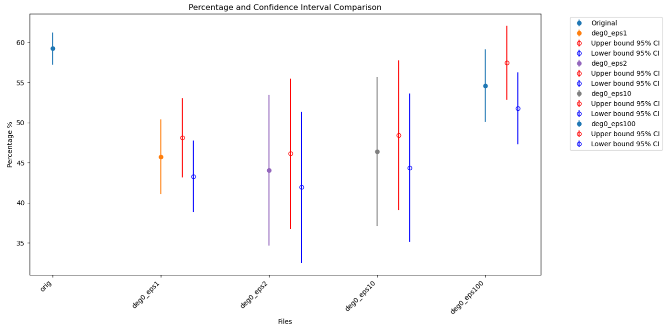

Fig. 7. Smoking quit attempt rates among current e-cigarette users, with 95% confidence intervals shown for both the original dataset and synthetic datasets generated using PrivBayes

Fig. 4-5 show the smoking quit attempt rates of former e-cigarette users (those who have stopped using) and current e-cigarette users with 95% confidence intervals on the original and synthetic datasets (generated using DataSynthesizer). We found that regardless of the combination of ϵ and k used to generate the dataset, the resulting quit attempt rates differ significantly from the original rates. Even when ϵ is increased to 100 (thereby adding very little noise), the discrepancy between the synthetic and original rates becomes larger. This indicates that the synthetic datasets (generated using DataSynthesizer) are unable to produce the correct statistical significance for quit attempt rates. The relationship between the increase in e-cigarette use and changes in smoking quit attempts does not exist in the synthetic datasets.

D. Comparing the PrivBayes with DataSynthesizer (the relationship between e-cigarette use and quit attempts)

Fig. 6-7 show the smoking quit attempt rates of former e-cigarette users and current e-cigarette users with 95% confidence intervals on the original and synthetic datasets (generated using PrivBayes). The general trend indicates that as ϵ increases, the rates of quit attempts in the synthetic datasets become closer to the original rate. When ϵ is large enough, the quit attempt rates in the synthetic datasets (generated using PrivBayes) are very close to the original rates. This indicates that the PrivBayes algorithm performs better than the DataSynthesizer algorithm in complex queries.

VI. Conclusion

The main observation is that, in general, complex queries exhibit significantly more variance compared to simpler queries and generally perform worse overall. The PrivBayes algorithm performs better than the DataSynthesizer algorithm in complex queries, while the DataSynthesizer algorithm excels in simple queries.

Increasing ϵ generally improves performance, which aligns with theory since increasing the privacy budget decreases the noise injected into the algorithm.

Higher-degree networks are generally preferred given a sufficient privacy budget, as they model dependencies better. This may be why the PrivBayes algorithm performs better than the DataSynthesizer algorithm in complex queries because the average number of parents chosen by PrivBayes increases with the set value of k.

VII. Future Research

The next step is to evaluate how DataSynthesizer and PrivBayes perform on other datasets and studies. We plan to implement this on the the Panel Study of Income Dynamics (PSID) dataset and compare the results. Additionally, we are interested in evaluating how data synthesizers perform on longitudinal studies, as they traditionally perform poorly in such contexts. We also plan to explore other data generators, such as McKenna, used in the NIST PSCR in 2019. This method involves more data preprocessing and post-processing, where relationships between attributes must be specified. We aim to use the Bayesian network information created by PrivBayes and DataSynthesizer to inform us of these relationships.

Acknowledgment

This work was supported by the funding from SUPER Research Experience The authors would like to thank Prof. Gong, Thomas Chen, and Project SUPER of Douglass Residential College

References

[1] Bowen, C. M., & Snoke, J. (2019). Comparative study of differentially private synthetic data algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge. *Unpublished Manuscript*.

[2] Zhu, S.-H., Zhuang, Y.-L., Wong, S., Cummins, S. E., & Tedeschi, G. J. (2017). E-cigarette use and associated changes in population smoking cessation: Evidence from US current population surveys. *BMJ*, *358*, j3262. https://doi.org/10.1136/bmj.j3262

[3] Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., & Xiao, X. (2017). PrivBayes. *ACM Transactions on Database Systems*, *42*(4), 1–41. https://doi.org/10.1145/3134428

[4] Ping, H., Stoyanovich, J., & Howe, B. (2017). DataSynthesizer. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. https://doi.org/10.1145/3085504.3091117

[5] Flood, S., King, M., Rodgers, R., Ruggles, S., Warren, J. R., Backman, D., Chen, A., Cooper, G., Richards, S., Schouweiler, M., & Westberry, M. (2023). *IPUMS CPS: Version 11.0 [dataset]*. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D030.V11.0