Journal of Privacy and Confidentiality (2010) **2**, Number 2, pp. 21–33

How Can We Analyze Differentially-Private Synthetic Datasets?

Anne-Sophie Charest*

Abstract. Synthetic datasets generated within the multiple imputation framework are now commonly used by statistical agencies to protect the confidentiality of their respondents. More recently, researchers have also proposed techniques to generate synthetic datasets which offer the formal guarantee of differential privacy. While combining rules were derived for the first type of synthetic datasets, little has been said on the analysis of differentially-private synthetic datasets generated with multiple imputations. In this paper, we show that we can not use the usual combining rules to analyze synthetic datasets which have been generated to achieve differential privacy. We consider specifically the case of generating synthetic count data with the beta-binomial synthetizer, and illustrate our discussion with simulation results. We also propose as a simple alternative a Bayesian model which models explicitly the mechanism for synthetic data generation.

Keywords: Synthetic datasets, Differential privacy, Beta-binomial synthetizer

1 Introduction

Statisticians working in data collection are faced with two conflicting objectives. On the one hand, their job is to collect and publish useful datasets for analysts to use to design public policies and build scientific theories. On the other hand, they must protect the confidentiality of their respondents. Not only is this a legal requirement, as respondents are usually assured that their data will remain confidential and will be used only for "statistical purposes," but protecting the confidentiality of the respondents is also essential for statistical agencies to keep the trust of the population, ultimately leading to better response rates and data accuracy.

A method for statistical disclosure limitation which has gained popularity recently is to keep the real dataset confidential and create synthetic datasets for publication. It was first suggested by Rubin (1993) to generate synthetic datasets using the framework of multiple imputation by sampling from the posterior predictive distribution, with the argument that because the synthetic data do not correspond to any actual individual, they preserve the confidentiality of the respondents. Synthetic datasets were also discussed in Fienberg (1994), Fienberg et al. (1998), and Little (1993), among others.

The generation and analysis of synthetic datasets based on multiple imputations have been studied extensively (see e.g., Raghunathan et al. (2003), Reiter (2002), and Reiter

*Department of Statistics, Carnegie Mellon University, mailto:acharest@andrew.cmu.edu

(2003)), and we have some results about the accuracy of estimates from suchcitation synthetic datasets. However, it is not yet clear exactly what confidentiality guarantees this method offers.

One way of quantifying confidentiality protection is with the idea of differential privacy, a powerful criterion introduced by Dwork (2006). Differential privacy protects the confidentiality of individual respondents no matter what amount of external information may be available to an intruder—an attractive feature given the growing amount of information available on the internet which could be used for linkage and re-identification.

A recent paper introduced an algorithm to generate synthetic datasets which satisfied differential privacy for count data (Abowd and Vilhuber, 2008). While differentially-private algorithms had already been proposed to solve several statistical problems, from mean estimation to fitting a support vector machine (see Dwork (2008)), this new technique is of great importance for statistical agencies, who often wish to publish microdata for the analysts.

There is, however, not yet any literature on the analysis of synthetic datasets created to achieve differential privacy. In fact, in Machanavajjhala et al. (2008), the only example in the literature where differentially-private synthetic datasets are created for real data, it seems as if the sanitized dataset is to be analyzed as if it was the real dataset. There is also no indication as to whether information about the synthetic data generation was provided to the users.

This paper is intended as a first attempt to address the question of the analysis of synthetic datasets created to achieve differential privacy. We focus our discussion on the analysis of synthetic values for count data using the algorithm proposed in Abowd and Vilhuber (2008), and of which an adapted version was used in Machanavajjhala et al. (2008).

In Section 2, we define precisely the criterion of differential privacy and describe the creation of differentially-private synthetic datasets for count data. In Section 3, we present the usual rules for the analysis of synthetic datasets created with multiple imputations and conclude that they are not appropriate to analyze synthetic datasets that were created to achieve differential privacy. The main point is that these combining rules assume that the synthetic data are generated from a posterior predictive distribution that uses noninformative prior distributions, which is not true in this case. In Section 4, we present a simple Bayesian method to take into account the synthetic data generation process in the analysis, and we show that it performs better to analyze differentially-private synthetic datasets. Section 5 contains a brief discussion and ideas for future work.

2 Differential Privacy

2.1 Definition

Differential privacy protects the information of every individual in the database against an adversary with complete knowledge of the rest of the dataset. In fact, by making sure that the released data does not depend too much on the information from any one respondent, differential privacy guarantees respondents that an attacker will not learn much more about their personal information, whether or not they accept to join the dataset.

Formally, we say that a randomized function κ gives ϵ-differential privacy if and only if for all datasets B_1 and B_2 differing on at most one element, and for all $S \subseteq \text{range}(\kappa)$,

$$Pr[\kappa(B_1) \in S] \leq \exp(\epsilon) * Pr[\kappa(B_2) \in S] \tag{1}$$

with the assumption that the larger value is in the left. For multivariate datasets with n rows and p columns, differing by one element means that the two datasets are identical except for one of the n rows. Loosely speaking, differential privacy ensures that the released information would be similar enough for similar input datasets that very little information could be gained from the released data about specific entries in the real dataset.

The constant ϵ must be specified by the user, and controls the level of confidentiality guaranteed by the randomized function κ. We can more easily interpret $\exp(\epsilon)$, which controls the ratio of the probability of a certain outcome for two datasets differing by at most one element. Differential privacy can also be interpreted from a Bayesian perspective as controlling the ratio of posterior to prior distributions, as discussed in Abowd and Vilhuber (2008). Smaller values of ϵ indicate greater confidentiality protection, since an intruder observing a certain outcome would then have little information as to which dataset it was generated from. At this point, no real guidelines have been suggested for appropriate choices of ϵ. For the extreme choice of $\epsilon = 0$, the output of the randomized function κ would have the same distribution no matter the observed dataset.

2.2 Generation of Differentially-Private Synthetic Datasets

In the case of synthetic data generation, the randomized function κ takes as input the real dataset and generates a synthetic dataset to be released. We may want to release multiple synthetic datasets, say M of them, in which case we can ensure overall ϵ differential privacy by simply generating each synthetic dataset independently with ϵ/M differential privacy requirement.

We now present an algorithm to generate synthetic datasets which satisfy ϵ differential privacy. We consider a dataset of the form $X = (x_1, \ldots, x_n)$, where $x_i \in \{0, 1\}$ for $i = 1, \ldots, n$ are dichotomous variables. We assume a binomial likelihood for the data and can thus reduce the dataset to its sufficient statistic $x = \sum_{i=1}^{n} x_i$. To protect the confidentiality of the respondents, we want to publish an ϵ differentially-private

synthetic dataset $\tilde{x}$ instead of the collected data x.

The mechanism proposed by Abowd and Vilhuber (2008) is to sample

$$\begin{aligned} \tilde{p} &\sim \text{Beta}(\alpha_1 + x, \alpha_2 + n - x), \\ \tilde{x} &\sim \text{Binomial}(\tilde{n}, \tilde{p}). \end{aligned}$$

The synthetic dataset $\tilde{x}$ is the one which is released. Note that we may use this method to generate a dataset of a size $\tilde{n}$ different from that of the original dataset, for example, if we want to keep n confidential. If we want multiple synthetic datasets, we simply reiterate this process to obtain $\tilde{p}_m$ and $\tilde{x}_m$, for $m = 1, 2, \ldots, M$, where M is the number of synthetic datasets desired, usually chosen to be 5 or 10.

The parameters α_1, α_2 will be referred to as differential privacy parameters for the remainder of the paper. To obtain ϵ differential privacy, we must pick $\alpha_j \geq \frac{\tilde{n}}{\exp(\epsilon)-1}$ for $j = 1, 2$. As in Abowd and Vilhuber (2008), we will use $\alpha_1 = \alpha_2$, where α_1 is the minimum value which guarantees ϵ-differential privacy. It could make sense in some cases to choose α_1 and α_2 based on our prior distribution for p (see Section 4.3.1), but in general the analyst is not the same person as the one creating the synthetic dataset, so this would not be feasible. The differential privacy parameters however, cannot depend on the observed dataset.

We can interpret this synthetic data generation process as generating from a perturbed posterior predictive distribution. The perturbation consists of using an implicit prior distribution of Beta(α_1,α_2) instead of our actual prior for p. Choosing $\alpha_1 = \alpha_2$ implies that this perturbing prior is centered at 0.5, with a spread depending on the size of α_1.

3 Analysis with Combining Rules

3.1 Usual Rules for Completely Synthetic Datasets

The generation of differentially-private synthetic datasets described in Section 2.2 mimics the generation of synthetic datasets using the multiple imputation framework as proposed in Rubin (2003). It may thus seem appropriate to analyze such datasets using the multiple imputation framework. In this section, we present the combining rules used in the multiple imputation framework and conclude, with theoretical arguments and simulations, that they are not appropriate for differentially-private synthetic datasets based on multiple imputations.

Suppose we are generating M completely synthetic datasets D_m, $m = 1, \ldots, M$, and we want to estimate one parameter of interest Q. We obtain from each of the datasets

an estimate q_m of Q and an estimate v_m of the variance of this estimator. Now, define

$$\begin{aligned} \overline{q}_M &= \frac{1}{M}\sum_m q_m, \\ \overline{v}_M &= \frac{1}{M}\sum_m v_m, \\ b_M &= \frac{1}{M-1}\sum_m (q_m - \overline{q}_M)^2. \end{aligned}$$

Rubin (1987) shows that when multiple imputations are used to correct for nonresponse, we should estimate the parameter of interest by $\overline{q}_M$ and the variance of this estimator by

$$T = (1 + 1/M)b_M + \overline{v}_M. \tag{2}$$

The variance estimator takes into account the variability of the data, the variance due to using only a finite number of imputations, and the randomness of the nonresponse mechanism. When synthetic datasets are generated for confidentiality purposes, the analyst controls the selection mechanism so there is no variability due to the nonresponse mechanism. Hence, the estimator must be modified. Raghunathan et al. (2003) derived

$$T_M = (1 + 1/M)b_M - \overline{v}_M \tag{3}$$

to estimate the variance of $\overline{q}_M$ in the context where multiple completely synthetic datasets are created for confidentiality purposes. For a great discussion of the difference between T and T_M, see Reiter and Raghunathan (2007). Confidence intervals can then be obtained based on t-distributions with degrees of freedom $\nu_M = (m-1)(1 - r_m^{-1})^2$, where $r_m = (1 + 1/M)b_M/\overline{v}_M$. However, the variance estimator T_M may be negative, so Reiter (2002) proposes the following, which is always positive:

$$T_M^* = max(0, T_M) + \frac{\tilde{n}}{n}\overline{v}_M I[T_M < 0], \tag{4}$$

where $\tilde{n}$ is the sample size for the synthetic datasets.

Note that in the special case that we are considering, we have $x \sim \text{Binomial}(n, p)$, so that the parameter of interest is $Q = p$, and our individual estimates are $q_m = \tilde{x}/\tilde{n}$ and $v_m = q_m * (1 - q_m)/\tilde{n}$.

3.2 Bias of $\overline{q}_M$

We already noted that to generate the synthetic datasets we used a perturbed version of the posterior predictive distribution. Recall that we add a prior distribution centered at 0.5, and whose implied prior sample size may be large with respect to the size of the observed data. We would then expect the synthetic datasets to yield sample estimates larger than (smaller than) the estimates from the real dataset if the estimate from the

real dataset is smaller than (larger than) 0.5, inducing bias in the combined estimate $\overline{q}_M$. We will show that this is indeed the case.

Let $\hat{p}_x = \frac{x}{n}$ be the estimator of p computed from the real dataset x. We want to compare this estimator to the one obtained from synthetic datasets generated given x, so that we compute $E[q_m|x]$, where the expectation is taken with respect to the randomness induced by the synthetic data generation. By the linearity of expectation, and the fact that q_m and $q_{m'}$ are identically distributed for $m, m' \in \{1, \ldots, M\}$, we have that $E[\overline{q}_M] = E[q_m]$. Thus, we only need to consider the case of a single synthetic dataset. We find that

$$\begin{aligned} E[q_m|x] &= E\left[\frac{\tilde{x}}{\tilde{n}}\middle| x\right] \\ &= \frac{1}{\tilde{n}} E\left[E\left[\tilde{x}|\tilde{p}\right]|x\right] \\ &= \frac{1}{\tilde{n}} E\left[\tilde{n}\tilde{p}|x\right] \\ &= \frac{\alpha_1 + x}{\alpha_1 + \alpha_2 + n}. \end{aligned}$$

Since we must have $\alpha_1 + \alpha_2 > 0$ to achieve differential privacy, the synthetic estimator is not unbiased for $\hat{p}_x$, the estimate obtained from the real dataset, for any fixed dataset.

What if we suppose a prior distribution $p \sim \text{Beta}(\gamma_1, \gamma_2)$ and average over all possible datasets? We then find that

$$E(\hat{p}_x) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}E[E(x|p)] = \frac{1}{n}E(np) = \frac{\gamma_1}{\gamma_1 + \gamma_2},$$

but

$$\begin{aligned} E(q_m) &= E\left(\frac{\alpha_1 + x}{\alpha_1 + \alpha_2 + n}\right) \\ &= E\left[E\left(\frac{\alpha_1 + x}{\alpha_1 + \alpha_2 + n}\middle| p\right)\right] \\ &= \frac{\alpha_1 + n\frac{\gamma_1}{\gamma_1 + \gamma_2}}{\alpha_1 + \alpha_2 + n}. \end{aligned}$$

The estimator from the real dataset and from the synthetic dataset both have the same expectation only in the case that $\alpha_1 = k\gamma_1$ and $\alpha_2 = k\gamma_2$, for some constant k. In other words, if the differential privacy parameters required to obtain ϵ-differential privacy correspond to the parameters for our prior distribution on p, then both estimators have the same expectation with respect to the distribution of the data and the distribution induced by the randomness in the synthetic data creation. But, the differential privacy parameters are not meant to represent our belief about the parameter for data generation and would most likely not be equal to the parameters in our prior for p.

Note that the bias of $\overline{q}_M$ depends on the parameters α_1 and α_2, which in turn depend on n and ϵ. As the difference between $\frac{\alpha_1}{\alpha_1+\alpha_2}$ and $\frac{\gamma_1}{\gamma_1+\gamma_2}$ increases, so does the bias of $\overline{q}_M$. Choosing $\alpha_1 = \alpha_2$ in our algorithm, $\overline{q}_M$ would be unbiased for p only when $p = 0.5$. We note that the bias will not asymptotically decrease to zero as $n \to \infty$ or $M \to \infty$ because in both cases α_1 and α_2 will increase with the same order of magnitude.

We now show results from a simulation where we fixed the true parameter p, created a true dataset of size n=100, generated M synthetic datasets of size $\tilde{n} = 100$ such that we had overall ϵ differential privacy, and computed $\overline{q}_M$. This process was repeated 100,000 times, and Table 1 shows the empirical relative bias (in %) of the estimates obtained.

Table 1: Relative bias (in %) of $\overline{q}_M$ as an estimator of p (based on 100,000 simulation runs)

ϵ	p	True Dataset	M=1	$M = 2$	$M = 5$	$M = 10$
2	0.25	0.12	23.88	53.84	80.30	90.05
2	0.50	0.05	0.05	-0.03	0.03	-0.00
250	0.25	0.01	0.05	-0.04	0.00	0.05

As predicted above, $\overline{q}_M$ is unbiased in the case where $p = 0.5$. When $p = 0.25$, the estimator is biased no matter how many synthetic datasets we use, with a bias reaching 90% when $M = 10$, for a reasonable requirement of $\epsilon = 2$. There is nothing particular about our choice of $p = 0.25$; similar results are seen for other values of p not equal to 0.5, with worse biases as p becomes more extreme. Note that the bias increases with the number of synthetic datasets. This is because as M increases we must use a smaller value of ϵ for each individual dataset that we create.

The estimator is also unbiased if $\epsilon = 250$ since with such a large ϵ the differential privacy parameters are practically both zero. There is a clear trade-off between the accuracy of the estimates and the confidentiality guarantees one can make.

3.3 Variance Estimation

We consider T_M and T_M^* as estimators of the variance of the estimator $\overline{q}_M$. An important assumption for the derivation of these rules is that the synthetic datasets are generated from the posterior predictive distribution based on noninformative prior distributions. This raises concerns about the validity of T_M and T_M^* to estimate the variance of $\overline{q}_M$, which are confirmed in the simulation presented below.

For this simulation, the conditions are the same as when we studied the bias, except that we also estimate the variance of $\overline{q}_M$. Table 2 shows the relative bias (in %) of the estimators T_M and T_M^*, where the true variance was also estimated from the simulation.

Table 2: Relative bias (in %) of T_M and T_M^* as estimators of the variance of $\overline{q}_M$. The bias of T_M is smaller than that of T_M^* for $\epsilon = 2$, but T_M often takes on negative values. (Based on 100,000 simulation runs.)

p	ϵ	M	Variance of $\overline{q}_M$ (x 10^{-4})	Relative bias of T_M (%)	Relative bias of T_M^* (%)	Negative T_M values (%)
0.25	2	2	21.10	34.28	127.21	49
0.25	2	5	6.68	64.60	273.57	40
0.25	2	10	2.92	86.74	526.68	40
0.50	2	2	23.62	27.34	112.96	48
0.50	2	5	6.95	60.68	262.12	40
0.50	2	10	3.02	79.85	507.31	40
0.25	250	2	37.10	0.20	36.26	44
0.25	250	5	25.96	0.35	19.92	20
0.25	250	10	22.26	0.18	9.48	9

When $\epsilon = 2$, T_M and T_M^* both overestimate the variance. This holds whether $p = 0.25$ or $p = 0.5$; there is nothing particular about $p = 0.5$ for the variance estimation. The relative bias of T_M is smaller than that of T_M^*, but T_M is negative for almost half of the runs which makes it hard to use for inferential purposes. As with the estimation of $\overline{q}_M$, the bias of T_M^* increases as ϵ decreases and M increases. We can, however, see from the table that the actual variance of $\overline{q}_M$ decreases as M increases so that, ignoring the increase in bias of T_M as M increases, it would be advantageous to use more synthetic datasets rather than less if we could correctly estimate the variance of our estimator. We note that in all cases we obtain unbiased estimates of $\overline{q}_M$ and of its variance if we use the real dataset.

3.4 Coverage Analysis

One could argue that it is relatively unimportant that $\overline{q}_M$ be unbiased and its variance be correctly estimated as long as confidence intervals obtained for the parameter of interest have nominal coverage. We conducted a small simulation study to look at the coverage of intervals created from $\overline{q}_M$ and T_M^* under the same conditions as in the previous example. Figure 1 shows the estimated coverage probabilities from 1000 repetitions.

We see that if $p = 0.5$, the overestimation of the variance of the estimator leads to coverages of almost 100%. This comes at the cost of very large, and therefore uninformative, confidence intervals. For $p = 0.25$, the results are also poor. The coverage barely reaches 80% when ϵ is set to 3.0 for two synthetic datasets and decreases as the number of synthetic datasets increases. Note that the results with the true datasets are not included in the graph, but the coverage was very close to the desired 95% in all cases.

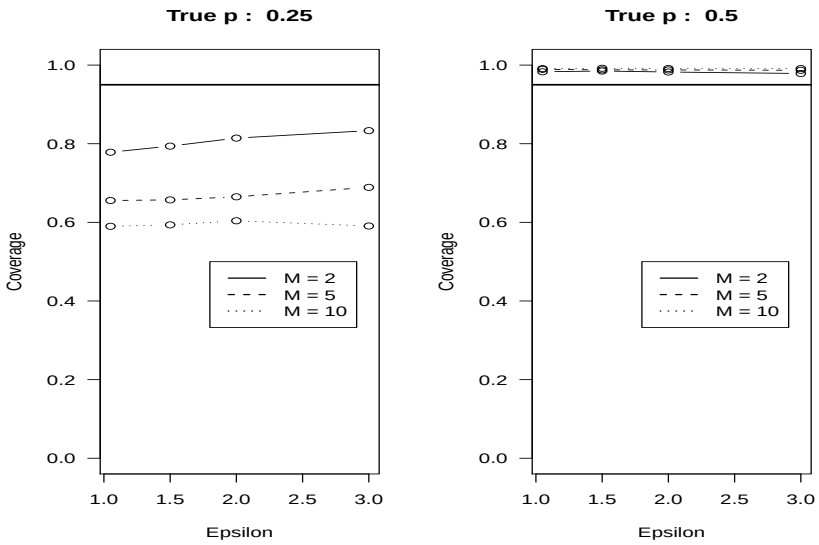

Figure 1: Coverage probabilities of 95% confidence intervals for p, using $\overline{q}_M$ and T_M^* (based on 10,000 iterations). For $p = 0.25$, the intervals do not achieve the nominal level and the coverage gets worse as M increases. When $p = 0.5$, increasing M leads to very high coverage, because the overestimation of the variance of the estimator creates very wide confidence intervals.

4 Analysis with Proposed Bayesian Model

We just showed that the combining rules which work very well when analyzing synthetic datasets generated from the posterior predictive distribution can not be applied if the synthetic dataset are generated to achieve differential privacy. One could try to derive new combining rules which take into account the differential privacy parameters. Instead, we model explicitly the data generation mechanism within a Bayesian model and conduct inference using the posterior distribution of p.

We use a conjugate prior for p, so that our complete model is:

$$\begin{aligned} p &\sim \text{Beta}(\gamma_1, \gamma_2), \\ x &\sim \text{Binomial}(n, p), \\ \tilde{p}_m &\sim \text{Beta}(\alpha_1 + x, \alpha_2 + n - x), \text{ for } m = 1, \ldots, M, \\ \tilde{x}_m &\sim \text{Binomial}(\tilde{n}, \tilde{p}_m), \text{ for } m = 1, \ldots, M. \end{aligned}$$

We only get to observe the vector $(\tilde{x}_1, \ldots, \tilde{x}_M)$ and the differential privacy parameters α_1, α_2, which we assume to be made available to the analyst. The posterior distribution for p does not have a closed form, but we can sample from it using MCMC. Updates

for p and $\{\tilde{p}_m\}_{m=1}^M$ are simple Gibbs updates:

$$p|\ x, \tilde{p}, \{\tilde{x}_m\}_{m=1}^M \sim \text{Beta}(\gamma_1 + x, \gamma_2 + n - x),$$
$$\tilde{p}_m | x, \tilde{x}_m, p \sim \text{Beta}(\alpha_1 + \tilde{x}_m + x, \alpha_2 + \tilde{m} - \tilde{x}_m + n - x) \quad \text{for } m = 1, \cdots, M.$$

To update x, one can use a Metropolis-Hastings step. At each iteration t, propose a new value x' to replace the current value x^t and accept a move from x^t to x' with probability $\min\left\{\frac{P(x')Q(x^t;x')}{P(x^t)Q(x';x^t)}, 1\right\}$, where

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \prod_{m=1}^{M} \frac{\tilde{p}_m^{\alpha 1 + x - 1}(1-\tilde{p}_m)^{\alpha 2 + n - x - 1}}{\Gamma(\alpha_1 + x)\Gamma(\alpha_2 + n - x)}$$

and $Q(x_1; x_2)$ is the probability that $x' = x_1$ given that $x^t = x_2$. One possible proposal distribution for x' is the following: let $x' = x^t + 1$ with probability k and $x' = x^t - 1$ with probability $1 - k$, unless $x^t = 0$ or $x^t = \tilde{n}$, in which case set $x' = 1$ and $x' = \tilde{n}$, respectively. The tuning constant k is chosen to achieve an acceptance rate of about 45%.

We can also fit this model more simply by using the JAGS software, which we did using the R package rjags. We run two independent chains and obtain a sample of 10,000 draws from each of them, keeping 1 in 20 observations after burn-in. Convergence is established using the Gelman-Rubin statistic.

We now present results to illustrate how this method performs compared to analyzing the true dataset directly. We will refer to the posterior distribution from the true dataset as the true posterior distribution, and the posterior distributions from the synthetic datasets as the synthetic posterior distributions. Ideally, the synthetic posterior distributions would be very close to the true posterior distribution.

In the example we consider, the true data set is $x = 30$ and $\epsilon = 2$. We use a uniform prior on $[0, 1]$ for p, so that the true posterior distribution is $\text{Beta}(31, 71)$, with expected value 0.3039 and variance 0.002053. Table 1 gives summary statistics for the synthetic posterior distributions obtained after generating $M = 1, 2, 5$ and 10 synthetic datasets.

Table 3: Mean and variance of the synthetic posterior distributions with comparison to the true posterior distribution. We vary the number of synthetic datasets but maintain overall 2-differential privacy. (Based on 1000 simulation runs.)

M	Posterior Mean	Relative bias of posterior mean (%)	Variance of the posterior distribution ($\times 10^{-3}$)
1	0.311	0.76	6.30
2	0.309	0.49	7.50
5	0.312	0.85	11.70
10	0.322	1.86	15.88

The synthetic posterior distributions are centered very close to the true posterior mean: modeling explicitly the synthetic data generation mechanism corrects for the bias it introduces, although maybe less so for larger M. The variance of the posterior distribution however increases greatly M increases. Recall that if we create M datasets and want an overall 2-differential privacy guarantee, we must create each dataset with $\epsilon = 2/M$. The added variability that comes from using a smaller value for ϵ is not offset by the increase in the number of datasets.

Note that the reason for the generation of multiple datasets was to estimate the variance of the estimator using the framework of multiple imputations. Since we incorporate the noise addition directly in our model, this is no longer necessary: we can obtain such an estimate directly from the MCMC output. The above results thus indicate that, in our simple setting and with n known, it would be optimal to create only one synthetic dataset.

Table 4 shows results of generating a single synthetic dataset for various values of ϵ. The bias of the synthetic posterior mean increases as the differential-privacy requirement becomes stronger. For $\epsilon < 0.5$, even incorporating the data generation process in the model is not sufficient to offset the bias introduced by the differential-privacy. As for the variance of the synthetic posterior distribution, it decreases as ϵ increases, but even for $\epsilon = 250$, which practically generates synthetic datasets from the true posterior distribution for x, the variance is almost tree times that of the true posterior distribution variance.

Table 4: Mean and variance of the synthetic posterior distributions with comparison to the true posterior distribution. We vary the value of ϵ and generate only one synthetic dataset. (Based on 1000 simulation runs.)

ϵ	Posterior Mean	Relative bias of posterior mean (%)	Variance of the posterior distribution ($\times 10^{-3}$)	Ratio to variance from true posterior
0.1	0.485	18.09	77.07	37.54
0.5	0.365	6.14	33.75	16.44
1	0.315	1.14	15.63	7.61
2	0.311	0.72	8.18	3.98
3	0.310	0.61	6.55	3.19
250	0.312	0.83	5.81	2.83

5 Discussion

In this paper our starting point was that statistical agencies may be interested in releasing synthetic datasets satisfying differential privacy. We showed that the combining rules used to analyze completely synthetic datasets generated with multiple imputations are not valid when the synthetic datasets have been generated to guarantee differential

privacy. We then proposed an inferential model which takes into account the data generation mechanism directly. We showed that this model allows for accurate estimation of p for moderately large values of ϵ, but with a larger associated variance.

It remains to decide whether or not the loss in efficiency associated with the differential privacy modeling is offset by the increase in privacy. There are no guidelines for the choice of the value of ϵ. Given that the usual combining rules can not be applied in this case, we believe that this Bayesian analysis model should be used when studying the accuracy of differentially-private synthetic datasets, and are currently working towards this goal.

We admit that creating data based on a single count is a very simple special case of synthetic data generation. Our work can easily be adapted to a dataset consisting of a vector of counts. Another important assumption of our model is that the true sample size n and the differential privacy parameters are available to the analysts. Publishing this information does not impact the confidentiality guarantees for the respondents in the dataset. Still, some statistical agencies may want to keep n confidential and decide not to publish n nor the differential privacy parameters, which can be used to infer n. Our approach could be adapted to this case by adding prior distributions for n, α_1 and α_2 as another hierarchy to the model. Although this would allow valid Bayesian inference, our results may be strongly dependent on these prior.

Acknowledgments

The author thanks Dr. Steve Fienberg, Dr. Jerry Reiter, Dr. John Abowd and Dr. Krishnamurty Muralidhar for helpful discussions, as well as a referee and an associate editor for valuable comments and suggestions. This research was supported by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office and by the NSF Grant BCS0941518 to the Department of Statistics at Carnegie Mellon.

References

Abowd, J. and Vilhuber, L. (2008). How protective are synthetic data? In *Privacy in Statistical Databases*, vol. 5262 of *LNCS*. Springer-Verlag. 239–246. doi:10.1007/978-3-540-87471-3_20

Dwork, C. (2006). Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming*, vol. 4052 of *LNCS*. Springer-Verlag. 1–12. doi:10.1007/11787006_1

— (2008). Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, vol. 4978 of *LNCS*. Springer-Verlag. 1–19. doi:10.1007/978-3-540-79228-4_1

Fienberg, S. E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics*, 10(2):115–132.

Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Statistics*, 14(4):485–502.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9(2):407–426.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the Map. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. 277–286. doi:10.1109/ICDE.2008.4497436

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure Limitation. *Journal of Official Statistics*, 19(1):1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531–544.

— (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–189.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

— (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9:462–468.

— (2003). Discussion on multiple imputation. *International Statistical Review*, 71(3):619–625.

Sarathy, R. and Muralidhar, K. (2011). Some additional insights on applying differential privacy for numeric data. In *Privacy in Statistical Databases*, vol. 6344 of *LNCS*. Springer-Verlag. 210–219.