Thesis Proposal: Creation and Analysis of Differentially-Private Synthetic Datasets

Anne-Sophie Charest
Department of Statistics
Carnegie Mellon University
October 20, 2010

Thesis committee:

Dr. Steve Fienberg (chair)
Dr. Brian Junker
Dr. Jerry Reiter (Department of Statistics, Duke University)
Dr. Alessandro Rinaldo

Abstract

Statistical agencies are faced with two conflicting objectives: protecting the privacy of their respondents, and providing researchers and policy makers with useful data. There exists a large body of literature on Statistical Disclosure Limitation (SDL) techniques, describing and evaluating methods for statistical agencies to share collected information to users while satisfying their confidentiality requirements. One proposal is to create and publish synthetic datasets. While methods to create and analyze such datasets have been published, it is still unclear how much privacy protection they offer. In our work, we look at the criterion of differential privacy, which protects every record in a dataset even against an intruder with complete knowledge of all other entries. In this thesis, we plan to create general procedures to generate and analyze synthetic datasets satisfying differential privacy. We also want to clearly measure the trade-off between the loss of information that comes from using synthetic datasets and the privacy offered to the respondents.

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 privacy 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.

There are two possible breaches of privacy: identity disclosure and attribute disclosure. Identity disclosure occurs when a respondent can be identified from the data released. Attribute disclosure occurs when the data released allows an intruder to recover the characteristics of one of the respondents. The recovery need not be exact, and indeed, following Delanius (1977), one may say that disclosure happens as soon as the release of the data increases the accuracy with which the value of the confidential answer of a respondent can be inferred. Although attribute disclosure is often associated with identity disclosure, identity disclosure by itself may violate the confidentiality agreement between the statistical agency and the respondent.

There exists a large body of literature on Statistical Disclosure Limitation (SDL) techniques, describing and evaluating methods for statistical agencies to share collected information to users while satisfying their confidentiality requirements. These include methods for data access restriction, data reduction, and data perturbation. Most are based on the intuitive idea that providing less data and providing perturbed data both reduce the likelihood of disclosure. While proposed SDL methods are evaluated to ensure that they allow analysts to obtain accurate statistical information, few rigorously quantify the extent to which they protect against identity and attribute disclosure.

It is also of importance for a good SDL method to be as transparent as possible, as this will help in three ways. First, providing details about the SDL mechanisms will allow users to take them into account in the statistical analyses to obtain valid inferences, ideally similar to the ones which would have been possible with the real data. Second, describing the SDL methods may help to convince the respondents that their privacy is important and will be respected. Finally, if the description of the SDL methods makes it clear that no personal information can be learned by mining through the released datasets it should deter possible intruders from trying to learn private information.

In this thesis, we study the still largely open problem of developing SDL methods which provide strict confidentiality guarantees yet allow agencies to share statistical knowledge with data users in a transparent manner. In the following section, we briefly review the literature on SDL methods and measures of privacy and data utility. We then state more precisely the goals of our thesis in section 3, provide some preliminary work and results in section 4, and discuss the proposed work in section 5.

2 Literature Review

2.1 SDL Methods

Statistical agencies faced with the need to protect the confidentiality of their respondents may use one of two approaches: restrict access to the sensitive dataset to certain trusted users, or release a restricted version of the dataset. An argument can be made that government statistics should be available to as many people as possible (see Fienberg (2001) and references therein), and thus our attention will be mainly on the second approach.

Statistical datasets can be restricted in two ways: via data reduction or data perturbation. An common method for data reduction is releasing only a subsample of the collected data, used for example by IPUMS-international (see McCaa et al. (2006)). Data reduction methods also include cell collapsing and cell suppression for contingency tables, and de-identification of individuals through the removal of identifiers.

There is also a wide array of data perturbation methods, including swapping, random noise addition, top-bottom coding and data shuffling. A common problem of most of these methods is that they are difficult for the analysts to take into account when making inferences. This is in part because details about the SDL procedures are not always made public, and in part because several users do not have the necessary statistical sophistication.

The most extreme type of data perturbation is the generation of synthetic datasets. It was first suggested by Rubin (2003) 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 necessarily preserve the confidentiality of the respondents. The generation and analysis of such synthetic datasets has been studied extensively (see for e.g. Raghunathan et al. (2003), Reiter (2002a) and Reiter (2003)), and we have some results about the accuracy of estimates from synthetic data sets, but it is not yet clear exactly what privacy guarantees they give.

Certain statistical agencies are also working on remote analysis servers to release sanitized results of user-specified statistical analyses. Proposals to lower the disclosure risk include subsampling, suppressing some of the output, and providing only approximate p-values (Sparks et al. (2008)). A difficulty with such endeavors, besides trying to accommodate all analyses that people may be interested in, is to provide privacy-preserving tools for exploratory analyses and model diagnostics. Also, systems based on queries to a remote server are vulnerable to combination attacks, in which different users combine information obtained during interactions with the remote server to reveal sensitive information.

2.2 Privacy Guarantees

To evaluate the success of SDL methods, we must define clearly what we mean by protecting the privacy of the respondents. Within the statistical literature, the focuss has been on estimating the probability of re-identification of respondents in released datasets. A general framework for computing such probabilities is laid out in Duncan and Lambert (1989) and Lambert (1993), and a detailed example is given in Reiter (2005). Other approaches are described in Fienberg and Makov (1998) and Forster and Webb (2007).

Some measures of privacy have also been developed especially for synthetic datasets. These include the median synthetic cell count versus the real cell count (Young et al. (2009)), which is based on the idea that an intruder will have difficulty predicting the number of individuals in the population belonging to a certain subgroup if the real counts are variable enough given a synthetic cell count. Another measure sometimes considered is the percentage of cells with a real count of zero among synthetic cells with a count of zero, which measures the risk of inferring population uniqueness based on sample uniqueness.

Overall, in the statistical literature most works focus on the probability of identity disclosure. Little has been done to quantify the attribute disclosure risk not resulting from identity disclosure.

Researchers from computer science approached the problem differently, developing theoretical privacy criteria and methods for creating publishable datasets with these properties. A first suggestion was k-anonymity (Sweeney (2002)). It is based on the division of records in equivalence classes by cross-classifying on quasi-identifiers, such as sex, age and zip code. A dataset will satisfy k-anonymity if each equivalence class contains at least k records. The idea

is that an attacker who knows the quasi-identifiers for an individual will not be able to identify the target with probability greater than $1/k$. While this definition reduces the risk of identity disclosure, there are still possibilities for attribute disclosure, for example if all records in an equivalence class have the same value for a sensitive attribute (see Li et al. and references therein for a discussion of this and other attacks). To preclude this attack, Machanavajjhala et al. (2008) introduced l-diversity, a criteria requiring that each equivalence class has at least l "well-represented" values for each sensitive attribut, where "well-represented" may be defined in several ways. This new definition also had problems and t-closeness was proposed as an alternative (Li et al. (2007)). It is a little more general, as it requires that the distribution of any sensitive attribute in an equivalence class be similar to the distribution of the attribute in the overall table.

All these criteria to control the coarseness of the published databases suffer from two common problems: they are sensitive to outside knowledge from the attacker, and they rely on a division of the measured variables into identifiers, quasi-identifiers and sensitive attributes.

A strongest definition was proposed a few years ago (Dwork (2006)) that does not assume anything about the intruder nor tries to differentiate between variables which should or should not be protected. By ensuring that the released statistical information does not depend too much on any particular record, differential privacy, as it is called, in fact protects the confidentiality of every individual response no matter what amount of external information may be available to an intruder. This is an attractive protection given the growing amount of information available on the internet which could be used for linkage and re-identification. Differentially private algorithms have already been proposed to solve several statistical problems, from mean estimation to fitting a Support Vector Machine (see Dwork (2008)).

We must also point out a recent effort by Kifer and Lin (2010) to axiomatize the idea of privacy. Their work provide new perspectives on the construction of privacy definition and mechanisms to obtain such data.

2.3 Utility Assessment

Disclosure risk can only be eliminated if the data are not released at all (Fienberg (2001)). As publication involves a certain risk, it should be weighted against the utility of releasing the data.

Some measures of data utility compare the inferences from the real dataset to that obtained with the data to be released. One can calculate for example the differences between point estimates or measure the overlap between intervals obtained using the real and the pertubed data. More general measures are based on how well one can discriminate between the original and the sanitized data (Woo et al. (2009)), as well as information-theoretic measures of the difference between the two distributions of the data (Domingo-Ferrer and Rebollo-Monedero (2009)).

Several frameworks have also been proposed to study the privacy-utility tradeoff (Duncan et al. (2003); Karr et al. (2006); Reiter (2009)). The idea is to pick, for a fixed level of privacy required, the SDL method which provides the largest utility. Guidelines for the choice of that privacy level however remain to be established.

Interestingly, the utility is not usually measured in terms of marginal utility compared to not releasing the data at all, but usually in terms of relative loss in utility incured by the privacy mechanism. This is implicitly assuming that all information should be published unless it can not be, probably a fair statement if that data was collected for a specific purpose, but maybe not adequate for other databases, such as query logs, or credit card transaction records.

3 Thesis Goals

Because of its versatility and independence on assumptions about the attacker, we consider differential privacy to be a good guide for a privacy requirement. Thus, in this thesis, we will study the generation and analysis of synthetic datasets with differential privacy guarantees. We have three main goals:

- Give methods for the generation of synthetic datasets with formal privacy guarantees, such as differential privacy
- Measure the trade-off between the loss of information that comes from using differentially private synthetic datasets and the privacy offered to the respondents
- Provide techniques for inference with the differentially private synthetic datasets

4 Demonstration of competence

So far, we have obtained results about the analysis of differentially private synthetic datasets for count data. We first define rigorously differential privacy and describe a method proposed by Abowd and Vilhuber (2008) to generate differentially private synthetic datasets for count data. We then present our results on the analysis of such datasets. We admit that creating a single count data is a very simple special case of synthetic data generation. We choose to use it here nonetheless for two reasons. First, this beta-binomial synthesizer is at the basis of the algorithm used by the U.S. Census Bureau to generate synthetic data for commuting patterns (see Machanavajjhala et al. (2008)), the only instance we know of where synthetic datasets were created with the constraint of achieving some form of differential privacy. Second, the simplicity of this setup allows us to very clearly illustrate the different problems with the usual inference method for synthetic datasets.

4.1 Differential Privacy

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. Loosely speaking, differential privacy ensures that the released information would be similar enough for similar input datasets that very little information can 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 privacy guaranteed by the randomized function κ. We can interpret more easily $\exp(\epsilon)$, 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 privacy 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.

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.

4.2 Synthetic Data Generation

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 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 data set $\tilde{x}$ 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 $\tilde{p}_i$ and $\tilde{x}_i$, for $i = 1, 2, \ldots, M$, 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 remaining of the paper. To obtain ϵ differential privacy, we must pick $\alpha_i \geq \frac{\tilde{n}}{\exp(\epsilon)-1}$. 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 Sect. 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 can however not 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.

4.3 Analysis with Combining Rules

The generation of differentially-private synthetic datasets described in the previous section mimics the generation of synthetic datasets using the multiple imputation framework, which was discussed in section 2. 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 imputateions are used to correct for nonresponse we should estimate the parameter of interest by $\overline{q}_M$ and the variance of this estimator by

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

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 privacy purpose, the analyst controls the selection mechanism, so there is no variability due to the nonresponse mechanism and the estimator must be modified. Reiter (2002b) 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 privacy purpose. For a great discussion of the difference between T_m 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 $r_m = (1 + 1/M)b_M/\overline{v}_M$. The variance estimator T_M may however be negative, so Reiter (2002b) proposes the following, which is always positive

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

where n_{syn} is the sample size for the synthetic datasets.

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

4.3.1 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 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 data set 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

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

The synthetic estimator is therefore not unbiased for the estimate obtained from the real data set, for any fixed data set.

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 data set and the synthetic data set estimator 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 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 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. The largest the difference is between $\frac{\alpha_1}{\alpha_1+\alpha_2}$ and $\frac{\gamma_1}{\gamma_1+\gamma_2}$ is, the largest the bias. Choosing $\alpha_1 = \alpha_2$ in our algorithm, $\overline{q}_M$ would we 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, then created a true dataset, generated M synthetic datasets such that we had overall ϵ differential privacy, and computed $\overline{q}_M$. This process was repeated 1000 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 1000 simulation runs)

ϵ	p	True Dataset	M=1	M = 2	M = 5	M = 10
2	0.25	0.16	6.00	13.50	20.57	22.55
2	0.50	0.07	0.01	-0.04	-0.07	-0.00
250	0.25	0.11	0.41	-0.31	-0.50	0.14

As predicted above, $\overline{q}_M$ is unbiased only in the case where $p = 0.5$. In the case where $p = 0.25$, the estimator is biased no matter how many synthetic datasets we use, with biases reaching 23% when using $M = 10$, for a reasonable requirement of $\epsilon = 2$. Using a larger value of ϵ, in our case the extreme value of 250, significantly reduces the bias of the estimator. There is a clear trade-off between the accuracy of the estimates and the privacy guarantees one can make. We note that 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 the more extreme p is.

Note that the bias actually 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. This is not a very desirable, and somewhat counterintuitive, feature. The method we propose in section 4.4 to analyze these datasets is better behaved: in this case the accuracy increases as M increase.

4.3.2 Variance Estimation

We consider T_M^* as an estimator of the variance of the estimator $\overline{q}_M$. An important assumption for the derivation of this rule is that the synthetic datasets are generated from the posterior predictive distribution. In our case, this is only true when the differential privacy parameters are equal to the parameters for our prior on p. This raises concerns about the validity of T_p 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 estimator T_M^*, where the true variance was also estimated from the simulation.

Table 2: Relative bias (in %) of T_p as an estimator of the variance of $\overline{q}_M$ (based on 1000 simulation runs)

p	ϵ	M	Variance of $\overline{q}_M$ (x10^{-2})	Relative bias of T_p (%)
0.25	2	2	22.40	54.35
0.25	2	5	6.42	251.00
0.25	2	10	3.05	503.95
0.50	2	2	23.57	63.09
0.50	2	5	7.09	225.99
0.50	2	10	3.12	466.29
0.25	250	2	39.42	-14.66
0.25	250	5	30.35	-15.33
0.25	250	10	25.46	-16.71

As before, and for the same reasons, the bias increases as ϵ decreases and M increases. For the variance estimation, there is nothing particular about $p = 0.5$; in all cases, the variance is overestimated. 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 as M increases, it would be advantageous to use more synthetic datasets than less if we could estimate correctly 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.

4.3.3 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 before. Fig. 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 three 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.

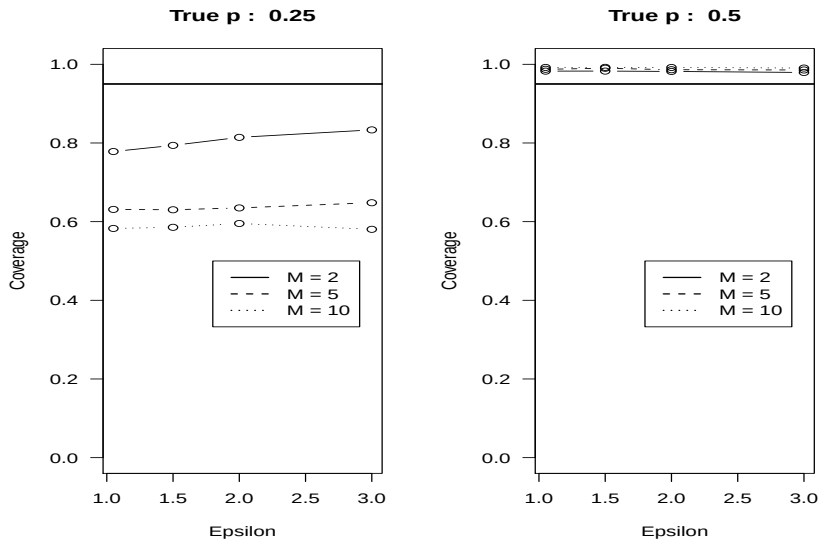

Figure 1: Coverage probabilities of 95% confidence intervals for p, using $\overline{q}_M$ and T_M^* (based on 10000 iterations). For $p = 0.25$, the intervals do not achieve the nominal level and the coverage gets worse as M increases. When $p = 0.5$, increasing M leads to very high coverage, because the overestimation of the variance of the estimator creates very wide confidence intervals.

4.4 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 data set are generated to achieve differential privacy. One could try to derive new combining rules which take into account the differential privacy parameters. Instead, we conduct inference using the posterior distribution of p, taking into account the synthetic data generation mechanism.

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

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

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 concept of differential privacy allows to release such information without increasing the concern for confidentiality breach. The posterior distribution for p does not have a closed form, but we can sample from it using MCMC. Updates for p and $\tilde{p}$ are simple Gibbs updates:

$$\begin{aligned} p|x, x_i, \tilde{p} &\sim \text{Beta}(\gamma_1 + x, \gamma_2 + n - x) \\ \tilde{p}_i|x, x_i, p &\sim \text{Beta}(\alpha_1 + \tilde{x}_i + x, \alpha_2 + \tilde{m} - \tilde{x}_i + n - x) \end{aligned}$$

whereas a Metropolis-Hastings step can be used to update x.

We now present results to illustrate how this method allows for appropriate analysis of differentially synthetic datasets. In this example, $n = 100$, $\epsilon = 2$, and we obtain 1000 draws from the posterior distribution using the JAGS software. Figure 2 compares our posterior

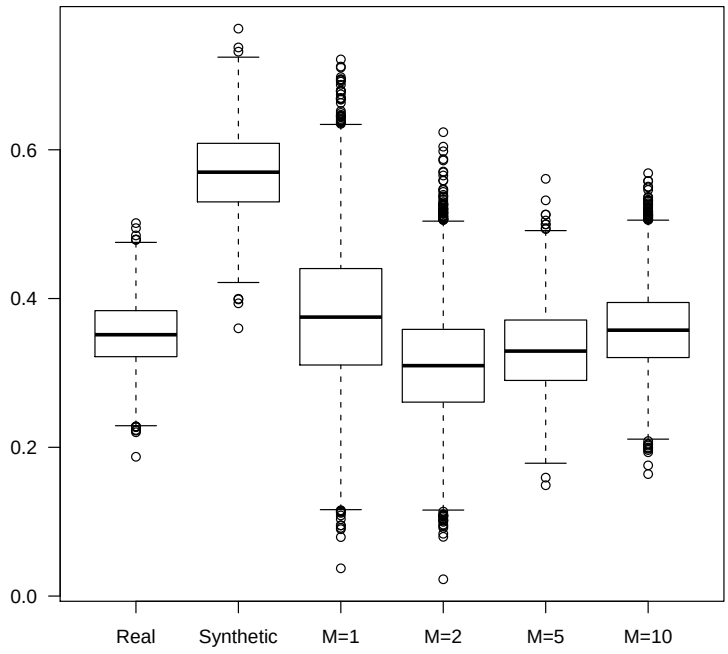

Figure 2: Posterior distributions for p. The left-most boxplot is the posterior when the real dataset is used in a simple beta-binomial model. To its right is the posterior obtained when analyzing a unique synthetic dataset as if it were real, which is in this case greatly shifted towards higher values of p. The last four boxplots show that when information about the synthetic data generation process is integrated in the model as described in section 4.4 combining several synthetic datasets allows to almost recover the posterior distribution that would have been obtained with the real dataset.

distribution for p to the one that would have been obtained using the real dataset and the one obtained from analyzing the synthetic datasets but ignoring the data generation mechanism.

We clearly see that using the synthetic data but ignoring the data generation mechanism (indicated by Synthetic in the graph) yields a posterior distribution for p much different than what would have been obtained with the real data set. It is centered at a much higher estimate for p. If the analysis specifically takes into account the synthetic data generation mechanism, we obtain posterior distributions which are much closer to the ones we would have obtained using the real data set. The similarity between the two posterior distribution increases as M increases, even though each data set then has to be created to satisfy ϵ/M-differential privacy. An analysis taking into account the SDL mechanism in the modeling phase thus appears promising to give protect data utility while having the rigorous guarantee of differential privacy.

5 Future work

Our goals for this thesis are to provide methods for the generation and the analysis of datasets satisfying differential privacy, measure the privacy-utility tradeoff of such methods, and provide guidelines to data disseminators. So far, we have showed that although the ideas from multiple imputation can be adapted to provide differentially private synthetic datasets for count data, the usual analysis methods for multiply imputed datasets are not valid in this case. We instead proposed a bayesian framework for modeling the data generation process explicitly during data analysis, and obtained promising results on its ability to maximize the utility of the synthetic datasets. We now discuss the proposed future work for the completion of the thesis.

5.1 Extend the data generation method to contingency tables

We provided a toy example where the dataset to protect was a single count. The method in Abowd and Vilhuber (2008) can actually be used for a vector of counts, by simply using a multinomial likelihood with a dirichlet prior. Similarly, our bayesian framework can easily be adapted to such datasets. Our next step if to find a method to create differentially-private synthetic contingency tables.

At this point, we believe that we can obtain differential privacy in this case in a similar fashion as for count data. The idea is to set up a bayesian model to describe the data, introduce additional noise by perturbing the prior distribution and then sample from the posterior predictive distribution. Although we could just consider a contingency table as a vector of counts and use the dirichlet-multinomial model, more complicated models are needed to preserve the relationships between the variables and protect the data utility.

Consider $x = (x_1, \cdots, x_n)$ a the vector of counts in a contingency table. We usually model $x_i \sim \text{Poisson}(\pi_i)$. One option is then to model the π_i following a log-linear model with various main effects and interactions. Such models have already been used in the context of statistical disclosure limitation to estimate probabilities of re-identification (e.g. in Fienberg and Makov (1998)). Skinner and Holmes (1998) propose a simpler version of the standard log-linear model by only allowing main effects in the model, but with an additional random effect. This is intended to protect the analyst from model mis-specification. Our goal is to show how to obtain differential privacy by sampling from the perturbed posterior predictive distribution for one of these models for contingency tables.

5.2 Develop methods for the analysis of differentially private synthetic datasets

With the generation of more complicated differentially private synthetic datasets, we must also provide methods for their analysis. It would be ideal to obtain some results on the possible accuracy of inferences made from differentially private synthetic datasets. Wasserman and Zhou (2010), for example, shows some asymptotic results for the convergence of empirical distributions of differentially-private synthetic datasets generated using two different methods. This paper shows that in general privacy schemes do not seem to yield minimax rates. We would like to analyze how our inference method compares.

Another aspect of interest when analyzing multiple synthetic datasets is that of model selection and tests of fit. Methods have been suggested for multiply imputed synthetic datasets (Kinney (2007)), but they must be adjusted to take the data generation method into account.

5.3 Apply our methods to real data

While we develop methods for the creation and analysis of differentially private synthetic datasets, we plan to apply these methods on at least one real dataset for which publication with confidentiality guarantees is necessary. The choice of this dataset will be done with our collaborators on a grant.

5.4 Consider the tradeoff between privacy and utility for the proposed methods

We have concentrated in this proposal on generating synthetic datasets satisfying the rigorous constraint of differential privacy. We assumed that the level of privacy desired, ϵ, was known to the data synthesizer. As part of our proposed work, we want to provide statistical agencies with guidelines for th choice of ϵ. This decision should be taken by considering the tradeoff between confidentiality and data utility of the synthetic datasets created.

We already discussed how to measure confidentiality using the definition of differential privacy. We intend to describe the parallels between this rigorous notion and other measures of confidentiality used by statistical agencies to test the confidentiality of synthetic datasets, as described in section 3. Ideally, our analysis would also provide the user with the corresponding data utility so that the privacy-utility tradeoff can be evaluated.

5.5 Further work

In addition to this planned proposed work, there are some more questions we are interested in looking at, time permitting.

- Consider how to operationalize the definition of differential privacy for continuous data. Some procedures have been developed to publish statistical summaries and inferences that satisfy differential privacy (McSherry and Talwar (2007); Dwork (2006)), but it is not clear how they generalize to publishing differentially-private synthetic datasets.
- Identify data perturbation/synthetic data creation techniques other than those based on multiple imputations that can be tweaked to give differential privacy.
- Think about other possible privacy guarantees that are a little less stringent than differential privacy, but still offer some formal privacy guarantees. An example which has already been proposed is probabilistic differential privacy (Machanavajjhala et al. (2008)). Instead of controlling the worst case probability ratio for all possible datasets, it requires only that differential privacy holds with large probability. Extensions of this sort may be necessary in practice since differential privacy is often hard to achieve.
- Extend these methods to account for practical difficulties with surveys, including complex sampling schemes, sampling weights (possibly with nonresponse or calibration adjustments), missing data and a very large number of variables, both categorical and continuous.

References

Abowd, J. and Vilhuber, L. (2008), "How Protective Are Synthetic Data?" in *Privacy in Statistical Databases*, pp. 239–246, Springer.

Delanius, T. (1977), "Towards a Methodology for Statistical Disclosure Control," *Statistik Tidskrift*, pp. 429–444.

Domingo-Ferrer, J. and Rebollo-Monedero, D. (2009), "Measuring Risk and Utility of Anonymized Data Using Information Theory," in *International Workshop on Privacy and Anonymity in the Information Society (PAIS)*, New York, USA, ACM Press.

Duncan, G. and Lambert, D. (1989), "Risk of Disclosure for Microdata," *Journal of Business and Economics Statistics*, 7, 207–217.

Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. (2003), "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map," .

Dwork, C. (2006), "Differential Privacy," in *The 33rd International Colloquium on Automata, Languages and Programming*, pp. 1–12, New York, ACM Press.

Dwork, C. (2008), *Differential Privacy: A Survey of Results*, vol. 112, pp. 1–19, Springer.

Fienberg, S. E. (2001), "Statistical Perspectives on Confidentiality and Data Access in Public Health," *Statistics in Medicine*, 20, 1347–1356.

Fienberg, S. E. and Makov, U. E. (1998), "Confidentiality , Uniqueness , and Disclosure Limitation for Categorical Data," *Journal of Official Statistics*, 14, 385–397.

Forster, J. J. and Webb, E. L. (2007), "Bayesian Disclosure Risk Assessment: Predicting Small Frequencies in Contingency Tables," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56, 551–570.

Karr, a. F., Kohnen, C. N., Oganian, a., Reiter, J. P., and Sanil, a. P. (2006), "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality," *The American Statistician*, 60, 224–232.

Kifer, D. and Lin, B.-R. (2010), "Towards an Axiomatization of Statistical Privacy and Utility," .

Kinney, S. K. (2007), "Model Selection and Multivariate Inference Using Data Multiply Imputed for Disclosure Limitation and Nonresponse," Ph.D. thesis, Duke University.

Lambert, D. (1993), "Measures of Disclosure Risk and Harm," *Journal of Official Statistics*.

Li, N., Li, T., and Venkatasubramanian, S. (2007), "t-Closeness : Privacy Beyond k-Anonymity and l-Diversity," in *Proceedings of the IEEE 23rd International Conference on Data Engineering*, no. 2.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008), "Privacy: Theory meets Practice on the Map," *2008 IEEE 24th International Conference on Data Engineering*, pp. 277–286.

McCaa, R., Ruggles, S., Davern, M., Swenson, T., and Palipudi, K. M. (2006), *IPUMS-International High Precision Population Census Microdata Samples : Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts 3 The Case for High Precision Samples : The USA Experience*, no. May.

McSherry, F. and Talwar, K. (2007), "Mechanism Design via Differential Privacy," *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003), "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics*, 19, 1–16.

Reiter, J. P. (2002a), "Satisfying Disclosure Restrictions with Synthetic Data Sets," *Journal of Official Statistics*, 18, 531–544.

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

Reiter, J. P. (2003), "Inference for Partially Synthetic, Public Use Microdata Sets," *Survey Methodology*, 29, 181–189.

Reiter, J. P. (2005), "Estimating Risks of Identification Disclosure in Microdata," *Journal of the American Statistical Association*, 100, 1103–1112.

Reiter, J. P. (2009), "Disclosure Risk and Data Utility for Partially Synthetic Data : An Empirical Study Using the German IAB Establishment Survey," *Journal of Official Statistics*, 25, 589–603.

Reiter, J. P. and Raghunathan, T. E. (2007), "The multiple Adaptations of Multiple Imputation," *Journal of the American Statistical Association*, 102, 1462–1471.

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

Rubin, D. B. (2003), "Discussion on Multiple Imputation," *International Statistical Review*, 71, 619–625.

Skinner, C. J. and Holmes, D. J. (1998), "Estimating the Re-identifi cation Risk Per Record in Microdata," *Journal of Official Statistics*, 14, 361–372.

Sparks, R., Carter, C., Donnelly, J. B., O'Keefe, C. M., Duncan, J., Keighley, T., and McAullay, D. (2008), "Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving Analytics." *Computer Methods and Programs in Biomedicine*, 91, 208–22.

Sweeney, L. (2002), "k-Anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10, 557–570.

Wasserman, L. and Zhou, S. (2010), "A Statistical Framework for Differential Privacy," *Journal of the American Statistical Association*, 105, 375–389.

Woo, M.-j., Reiter, J. P., Oganian, A., and Karr, A. F. (2009), "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation," *The Journal of Privacy and Confidentiality*, 1, 111–124.

Young, J., Graham, P., and Penny, R. (2009), "Using Bayesian Networks to Create Synthetic Data," *Journal of Official Statistics*, 25, 549–567.