Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data

David McClure*, **Jerome P. Reiter****

*Department of Statistical Science, Duke University, Box 90251, Durham, 27708, USA.

**Department of Statistical Science, Duke University, Box 90251, Durham, 27708, USA.

E-mail: david.mcclure52@gmail.com, jerry@stat.duke.edu

Abstract. We compare the disclosure risk criterion of ϵ-differential privacy with a criterion based on probabilities that intruders uncover actual values given the released data. To do so, we generate fully synthetic data that satisfy ϵ-differential privacy at different levels of ϵ, make assumptions about the information available to intruders, and compute posterior probabilities of uncovering true values. The simulation results suggest that the two paradigms are not easily reconciled, since differential privacy is agnostic to the specific values in the observed data whereas probabilistic disclosure risk measures depend greatly on them. The results also suggest, perhaps surprisingly, that probabilistic disclosure risk measures can be small even when ϵ is large. Motivated by these findings, we present an alternative disclosure risk assessment approach that integrates some of the strong confidentiality protection features in ϵ-differential privacy with the interpretability and data-specific nature of probabilistic disclosure risk measures.

Keywords. Bayesian, Confidentiality, Disclosure, Risk, Synthetic

1 Introduction

Organizations seeking to share microdata, i.e., data on individual records, with others are obligated to protect the confidentiality of data subjects' identities and sensitive attributes. Even when direct identifiers such as names and addresses have been removed, data subjects still might be at risk. For example, ill-intentioned users—henceforth called intruders—might be able to link records in the released data file to records in other databases (that contain direct identifiers) by matching on common characteristics. Hence, it is prudent for organizations to assess the risks of unintended disclosures of confidential information and to release data in ways that have acceptable risks (while being useful for broad classes of analyses).

Both the statistical science and computer science communities have developed a variety of criteria and methods for quantifying disclosure risks. In this article, we consider a criterion developed by the cryptography community, ϵ-differential privacy (Dwork, 2006). This criterion, which promises strong guarantees of confidentiality, has been a focus of intense research in both communities. For example, a variant of differential privacy was used to assess and ensure confidentiality protection in the U.S. Census Bureau's OnTheMap microdata product (Machanavajjhala *et al.*, 2008), which enables users to make maps of where people live and work. Other examples of research related to releasing microdata that satisfy ϵ-differential privacy include Blum *et al.* (2008), Barak *et al.* (2007), Abowd and Vilhuber (2008), and Charest (2010).

To our knowledge, there have been few numerical comparisons of ϵ-differential privacy and statistical disclosure risk measures; the most closely related works are by Abowd and Vilhuber (2008), Wasserman and Zhou (2010) and Sarathy and Muralidhar (2011). In this article, we explore relationships between ϵ-differential privacy and statistical disclosure risk measures for microdata, by which we mean measures based on probabilities that intruders could learn information about data subjects given the released data and a set of assumptions about the intruder's knowledge and behavior (e.g., Duncan and Lambert, 1986, 1989; Fienberg *et al.*, 1997; Reiter, 2005a; Reiter and Mitra, 2009). Using illustrative simulations, we reach several findings with relevance to disclosure risk assessment and data release. First, whereas differential privacy is agnostic to the specific values in the sampled data, probabilistic disclosure risk measures depend greatly on them. This difference makes a general mapping between the two paradigms elusive. Second, depending on the characteristics of the data, probabilistic disclosure risk measures can be small even when ϵ is relatively large. For example, in some simulation scenarios the probabilistic disclosure risk measures were small even when $\epsilon = 1000$, whereas in other scenarios the probabilistic disclosure risk measures were small only when $\epsilon < 10$. Additionally, in some scenarios decreasing ϵ from ten to one did not noticeably decrease the probabilistic disclosure risk measures (but significantly worsened the quality of the synthetic data). Motivated by these findings, we also present an alternative disclosure risk assessment approach that integrates some of the strong privacy protection in ϵ-differential privacy with the interpretability and data-specific nature of probabilistic disclosure risk measures.

The remainder of the article is organized as follows. In Section 2, we briefly review ϵ-differential privacy and probabilistic disclosure risk measures. In Section 3, we present the simulation design and results. In Section 4, we present the proposal for the alternative disclosure risk assessment approach. In Section 5, we conclude with a discussion of issues related to releasing differentially private microdata.

2 Disclosure Risk Measures

Throughout the text, we use Y_j to denote the data value for the jth record, where $j = 1, \ldots, n$. We let $D = (Y_1, \ldots, Y_n)$ be the collected data, which we assume to be complete. The goal is to release a version of D, which we call D^*, that has acceptable disclosure risks according to some criteria. For simplicity, we presume each Y_j is a scalar quantity, although the ideas here extend conceptually to multivariate data.

2.1 ϵ-Differential Privacy

A randomized function $f : D \to f(D)$ gives ϵ-level differential privacy for D if

1. $\forall$ possible data sets D_1,D_2 that differ on at most one element, and

2. $\forall S \subseteq \mathit{Range}(f(D))$,

we have

$$\frac{p(f(D_1) \in S)}{p(f(D_2) \in S)} \leq exp(\epsilon). \tag{1}$$

The definition is interpreted for microdata releases by considering f to be the data generator and $f(D)$ to be possible datasets that could result; we refer readers to Abowd and Vilhuber (2008) for details.

Differential privacy represents a strong guarantee of confidentiality. In a differentially private dataset, even an intruder with access to all but one record in D does not learn much (where much is governed by the value of ϵ) about that unknown record. The confidentiality guarantee holds regardless of the intruder's prior knowledge about D. Further, it holds for arbitrary D, in that ϵ-differential privacy is a property of the method used to create D^* rather than the actual values of D or D^*. When differential privacy is satisfied, the data disseminator need not hide information about the process used to generate D^*, except of course for actual values of D. As a negative, however, it can be difficult to ensure that non-trivial randomizers f satisfy the conditions of ϵ-differential privacy when releasing complex D^*.

2.2 Probabilistic Disclosure Risk Measure

We suppose that an intruder seeks to learn the value of Y_j for some record j in D. Let A represent the information known by the intruder about records in D. Let S represent any information known by the intruder about the process of generating D^*. For example, the data disseminator might tell the public that D^* is created by aggregating certain categories, or that values in D^* are simulated from particular statistical models. Given (D^*, A, S), the intruder's density for Y_j is

$$p(Y_j \mid D^*, A, S) \propto p(D^* \mid Y_j, A, S)p(Y_j \mid A, S). \tag{2}$$

Here, $p(Y_j \mid A, S)$ is the intruder's prior distribution on Y_j based on (A, S), and D^* serves to sharpen the intruder's prior beliefs about Y_j. We note that these probabilities are specific to the released data and original data; hence, they do not conform to the definition of differential privacy which considers all possible released and original datasets.

These probabilities can be manipulated to form a variety of disclosure risk measures. For example, a metric could be based on the distance between the posterior mode or expectation computed with (2) and the actual Y_j. The probabilities in (2) also can form the basis of identification disclosure risk measures (Skinner and Shlomo, 2008), e.g., by following the approach outlined in Drechsler and Reiter (2008).

Probabilistic risk measures like (2) have appealing features. They are readily interpretable and facilitate targeted disclosure protection strategies: records with high probabilities are at higher risks of disclosure. They enable investigation of risk under different scenarios for A and S, which helps data disseminators understand a full picture of potential risks and the impacts of disclosure treatment. Probabilistic disclosure risk measures can incorporate uncertainty due to random sampling of D (Reiter, 2005a; Drechsler and Reiter, 2008). However, because specification of intruder's knowledge is required, the probabilistic risk measures may not accurately portray risk when intruders know more than is specified in A. Additionally, it can be challenging to compute (2) with complex D^*.

To facilitate comparisons of ϵ-differential privacy and probabilistic measures, we set A and S to mirror differential privacy requirements as closely as possible. Specifically, we assume that the intruder knows exact values of all records except one record j; we label this dataset

D_{-j}. We also assume that the data disseminator releases all details about the nature of the data generation procedure; this is clarified further below in the context of synthetic data generation.

3 Simulation Studies

We compare ϵ-differential privacy and probabilistic measures using a simple but informative simulation scenario. For $j = 1, \ldots, n = 1000$, let Y_j be a draw from a Bernoulli random variable with probability p, and let $D = (Y_1, \ldots, Y_n)$. We consider four separate scenarios in which $p \in \{.001, .3, .5, .999\}$. We generate D^* based on the framework of fully synthetic data (Rubin, 1993; Raghunathan *et al.*, 2003; Reiter, 2005b; Reiter and Raghunathan, 2007; Drechsler *et al.*, 2008; Caiola and Reiter, 2010). To ease computation and since our primary focus is on risk comparisons, we let each D^* comprise a single rather than multiple synthetic datasets. With multiple datasets, data disseminators would have to factor in the additional information available to intruders in both the selection of ϵ and the computation of (2); see Reiter and Mitra (2009) for discussion of the latter.

3.1 Generating and Analyzing Differentially Private Synthetic Data

The synthetic data generation approach outlined in Raghunathan *et al.* (2003) does not necessarily result in D^* with a pre-specified level of ϵ-differential privacy. We therefore use a synthetic data generator akin to that of Abowd and Vilhuber (2008), which guarantees ϵ-differentially private synthetic data. For any D, we generate D^* by simulating $n_s = 1000$ new values of Y_j from their posterior predictive distribution based on a Beta($\alpha_\epsilon, \beta_\epsilon$) prior distribution for p, i.e.,

$$p(Y_j^* \mid D, \alpha_\epsilon, \beta_\epsilon) = \text{Bernoulli}\left(\frac{\sum_{j=1}^{n} Y_j + \alpha_\epsilon}{n + \alpha_\epsilon + \beta_\epsilon}\right). \tag{3}$$

Here, α_ϵ and β_ϵ do not represent prior beliefs about p as is usual in Bayesian modeling. Rather, they are set to make the synthesis process approximately invariant to values in D at a level implied by ϵ. As α_ϵ and β_ϵ get larger, the influence of the observed data on the posterior predictive distribution weakens, resulting in decreased ϵ levels. It can be shown that, for any $\epsilon > 0$ under the Bernoulli model, setting $\alpha_\epsilon = \beta_\epsilon = \frac{1}{e^{\epsilon/n_s} - 1}$ induces ϵ-differential privacy when releasing D^*; see Appendix A for details.

The analyst of D^* should account for the manner in which the data were generated when making inferences for parameters (p in these simulations). Here we do so via a Markov Chain Monte Carlo (MCMC) algorithm like the one developed by Charest (2010). This algorithm requires analysts to know the values of α_ϵ, β_ϵ, and n. We assume that these are part of S, along with the form of the posterior predictive distribution (but not $\sum_j Y_j$) in (3).

For any (D, D^*), we measure the utility of D^* with a criterion based on expected squared error loss,

$$U_{(D,D^*)} = \log_{10} \frac{E((p_0 - p)^2 \mid D^*)}{E((p_0 - p)^2 \mid D)}. \tag{4}$$

Here, p_0 is the true value of p per the simulation design, and the expectations are over the posterior distribution of p given the appropriate data; see Appendix B for computational details. The larger the value of $U_{(D,D^*)}$, the lower the quality of inferences about p_0 from D^*. We use log base 10 purely to facilitate graphical displays of the simulation results. As we shall see, the utility at small ϵ is significantly worse than the utility at large ϵ, even when viewed on the log scale.

3.2 Evaluations of Absolute Probabilistic Disclosure Risk

For each value of p_0 and each value of $\epsilon \in \{1000, 100, 10, 2, .2, .01\}$, we generate ten pairs of (D, D^*) to ensure that findings do not reflect unusual simulation events. For each (D, D^*), we compute the probabilistic disclosure risk, $R_j = Pr(Y_j = 1 \mid D^*, D_{-j}, S)$, for one record with true $Y_j = 1$. If D does not contain any records with $Y_j = 1$, we instead compute $Pr(Y_j = 0 \mid D^*, D_{-j}, S)$ for one record with $Y_j = 0$. For illustrative purposes, we assume that $R_j > .5$ represents an unacceptably high risk in absolute terms. Of course, some data disseminators might prefer lower absolute risk thresholds; the choice of threshold is not central to the findings in this subsection. For all scenarios, we set the intruder's prior distribution to $p(Y_j = 1 \mid D_{-j}, S) = .5$. Hence, values of $R_j \approx .5$ indicate that the intruder has learned little about Y_j from D^*. Setting $p(Y_j = 1 \mid D_{-j}, S) = w \neq .5$ results in similar interpretations centered around w rather than .5.

With this simulation design, R_j is computed as follows. Let $X = \sum_{j=1}^{n} Y_j$; let $X_{-j} = \sum_{k \in D_{-j}} Y_k$; and, let $X^* = \sum_{j=1}^{n} Y_j^*$ be the sum of the synthetic values. We have

$$\begin{aligned} R_j &= \frac{P(D^* \mid Y_j = 1, D_{-j}, \alpha_\epsilon, \beta_\epsilon)P(Y_j = 1 \mid D_{-j}, \alpha_\epsilon, \beta_\epsilon)}{\sum_{y=0}^{1} P(D^* \mid Y_j = y, D_{-j}, \alpha_\epsilon, \beta_\epsilon)P(Y_j = y \mid D_{-j}, \alpha_\epsilon, \beta_\epsilon)} \\ &= \frac{Bin(X^*, n, p = \frac{X_{-j}+1+\alpha_\epsilon}{n+\alpha_\epsilon+\beta_\epsilon})P(Y_j = 1)}{Bin(X^*, n, p = \frac{X_{-j}+1+\alpha_\epsilon}{n+\alpha_\epsilon+\beta_\epsilon})P(Y_j = 1) + Bin(X^*, n, p = \frac{X_{-j}+\alpha_\epsilon}{n+\alpha_\epsilon+\beta_\epsilon})P(Y_j = 0)}, \quad (5) \end{aligned}$$

where $Bin(a, b, c)$ is the probability of a for a binomial distribution with b trials and probability of success c, and where $P(Y_j = y) = P(Y_j = y \mid D_{-j}, \alpha_\epsilon, \beta_\epsilon)$. We use binomial probabilities since the specific ordering of the values in D^* does not carry any information. We note that intruders also could multiply the n probabilities as ordered in D^*. The result would be identical to (5), since the binomial coefficient cancels from (5).

Figure 1 displays plots of R_j versus U_{D,D^*} for the 60 datasets when $p_0 = .5$. Figure 2 displays the analogous plot when $p_0 = .001$. Taking both plots together, it is clear that the ϵ-level required to ensure $R_j \approx .5$ varies with the value of p_0. When $p_0 = .001$, setting $\epsilon = 1000$ results in several D^* with R_j near an unacceptably high .90, whereas setting $\epsilon < 10$ results in D^* with R_j in the vicinity of .5. When $p_0 = .5$, there is little incremental risk in these datasets even when $\epsilon = 1000$. Of course, for any D there could be some D^* for which the risks are unacceptably large; hence, it is important to compute R_j for any D^* under consideration for release. We note the variability in R_j for any given ϵ arises because the measure is specific to the generated (D, D^*), which differ across replications.

These results can be understood further as follows. As α and β get larger (ϵ decreases), they dominate X_{-j} and n, so that p in (5) goes to $\frac{\alpha_\epsilon}{\alpha_\epsilon+\beta_\epsilon}$. Thus, for small ϵ, for any X^* (and implicitly p_0) we have $Bin(X^*, n, p) \approx Bin(X^*, n, \frac{\alpha_\epsilon}{\alpha_\epsilon+\beta_\epsilon})$. As a result, the binomial likelihoods in (5) cancel, and we are left with $R_j = \frac{P(Y_j=1)}{P(Y_j=1)+P(Y_j=0)} = P(Y_j = 1)$. Thus R_j converges to the intruder's prior probability, so that the synthetic data tell the intruder almost nothing about the real data, as is required in differential privacy. Since we set $P(Y_j = 1) = .5$, R_j converges to .5 for small ϵ, regardless of p_0.

When p is near zero or one, the variance of the binomial distribution gets very small. Small changes to p in (5), which would arise when adding to X_{-j}, can cause significant changes

$p_0 \backslash \epsilon$	1000	100	10	2	.2	.01
.001	.125	.036	.0101	.00372	.000578	3.13e-05
.3	.00718	.00702	.00578	.00328	.000576	3.13e-05
.5	.00655	.00643	.00543	.00321	.000575	3.13e-05
.999	.0983	.0350	.0100	.00372	.000578	3.13e-05

Table 1: Expected increase in risk for various (p_0, ϵ) combinations using $p(Y_j|A, S) = .5$. Expected risk increases are small for p_0 near .5, even for large ϵ.

in the probabilities of generating particular synthetic datasets. For example, consider when $\epsilon = 1000$ and $\alpha_\epsilon = \beta_\epsilon = .58$. Suppose that $X_{-j} = 0$ and $Y_j = 1$, which easily could happen when $p_0 = .001$. We have $p(X^* = 3 \mid n = 1000, p = \frac{.58}{1001.16}) = .018$ and $p(X^* = 3 \mid n = 1000, p = \frac{1.58}{1001.16}) = .135$, which results in $R_j = \frac{.135}{.135+.018} = .88$. We observed this situation once in our simulations with $p_0 = .001$, which is apparent in Figure 2. However, when $p_0 = .5$ the mass of the binomial distribution is spread out, so that there is very little change in the probabilities of particular synthetic datasets when we change p in (5) by adding to X_{-j}. Hence, R_j is generally close to $P(Y_j = 1)$ when p_0 is near 0.5. This logic holds even if we use unequal values of $(\alpha_\epsilon, \beta_\epsilon)$ to engender ϵ-differential privacy with small ϵ.

To quantify these trends without relying only on a modest number of simulated datasets, we introduce the expected increase in risk for releasing D^*; that is, how much R_j increases over the prior probability that $Y_j = 1$. Specifically, for any data generation procedure and possible p_0, we compute

$$ER_j = \sum_{\text{All } X, X^*} (max(R_j, p(Y_j|A, S)) - p(Y_j|A, S))p(X^*, X|p_0). \quad (6)$$

Here, the use of the maximum discards cases where $R_j < p(Y_j|A, S)$, i.e., when the intruder's posterior belief about Y_j is less accurate than the prior belief. For our simulations, we have

$$p(X^*, X|p_0) = p(X^*|X, S)p(X|p_0) = Bin(X^*, n, p = \frac{X + \alpha_\epsilon}{n + \alpha_\epsilon + \beta_\epsilon}) * Bin(X, n, p = p_0). \quad (7)$$

This probability is easily computed: simply loop over all possible (X, X^*), computing (5) for each combination to obtain R_j. As before, for cases when $X = 0$, we set $Y_j = 0$ when computing R_j.

Table (1) displays ER_j for each (p_0, ϵ) combination, using $p(Y_j = 1|A, S) = 0.5$. The results confirm the trends in Figure 1 and Figure 2. Regardless of p_0, as ϵ decreases the quantities all converge to zero, since X (and therefore p_0) has almost no effect on X^* (and therefore R_j) for small ϵ. At $\epsilon = 1000$, we expect an increase in risk of less than .01 when $p_0 = 0.5$, but an increase of .125 when $p_0 = .001$.

These results, in addition to results based on other values of p_0 and w not displayed here, suggest that populations with p_0 near zero or one, i.e., some data values are rare, tend to produce scenarios with higher probabilistic disclosure risks. With small p_0, it is possible that a generated D contains only a single (or very few) observations with $Y = 1$. When ϵ is large and D^* contains at least one record with a simulated Y equal to one, the likelihood in (2) is much larger for $Y_j = 1$ than $Y_j = 0$, which results in a high R_j. A similar phenomenon

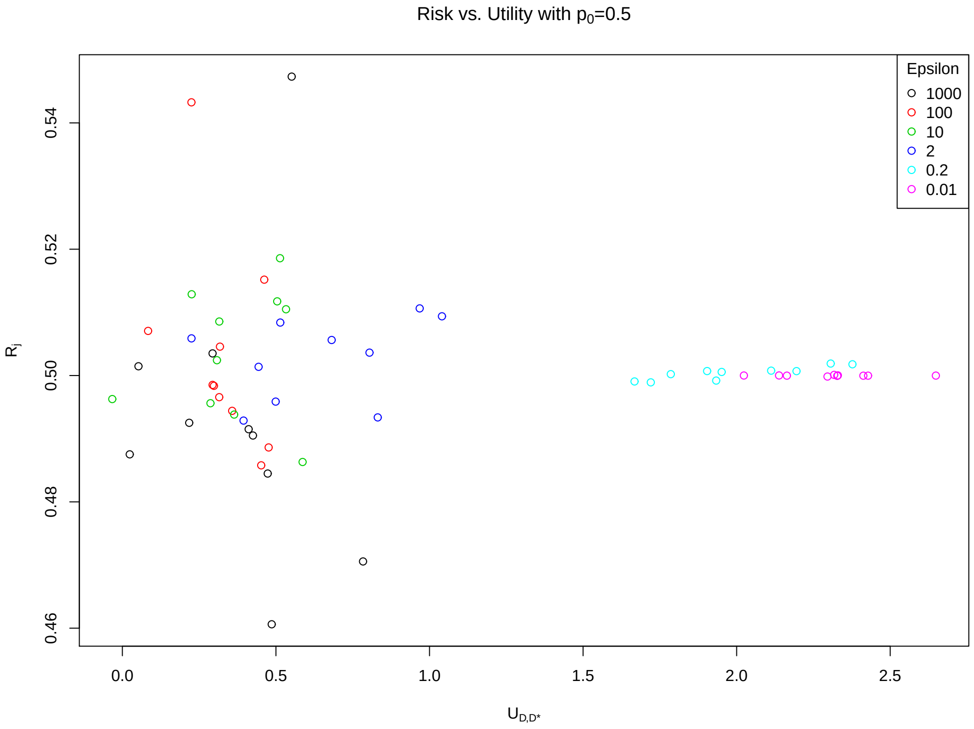

Figure 1: Absolute probabilistic disclosure risks R_j versus data utility $U_{(D,D^*)}$ when $p_0 = .5$. Each point represents a single (D, D^*) pair. Ten data sets are simulated for each ϵ level. Since $p(Y_j = 1 \mid D_{-j}, S) = .5$, $R_j \approx .5$ indicates the intruder learns very little from D^*.

occurs when $p_0 = .999$. This typically is not an issue when $p_0 = .5$, because here any particular Y_j does not have much influence on the posterior predictive distribution in (3).

In these simulations, setting ϵ to be very small seriously degrades inferences while providing only marginal reductions in probabilistic disclosure risks. For example, setting $\epsilon < 1$ resulted in basically the same values of R_j as when $\epsilon = 2$; but, setting $\epsilon < 1$ reduces $U_{(D,D^*)}$ significantly compared to $\epsilon = 2$, even when viewed on the log scale. The simulated values of $U_{(D,D^*)}$ range from 2 to 2.7 (roughly 100 to 500 before taking $\log_{10}$) when $\epsilon = .01$ and from .25 to 1.1 (roughly 2 to 13 before taking $\log_{10}$) when $\epsilon = 2$. Further, when $p_0 = .5$ the data disseminator that sets $\epsilon \geq 100$ arguably loses little in terms of probabilistic disclosure risk protection while gaining enormously in terms of data quality. Thus, the choice of ϵ has a strong influence on utility as well as confidentiality, and the degree of its influence depends on the nature of D. This suggests that data disseminators using differential privacy could benefit from evaluating trade offs between risk and utility of different ϵ when generating D^*. For example, in the simulation scenarios data disseminators could use larger ϵ for populations with $p_0 \approx .5$. We note that releasing $m > 1$ synthetic datasets such that the collection satisfies ϵ-differential privacy generally would exacerbate the problems with data utility, since the data disseminator would need to use ϵ/m when generating each dataset.

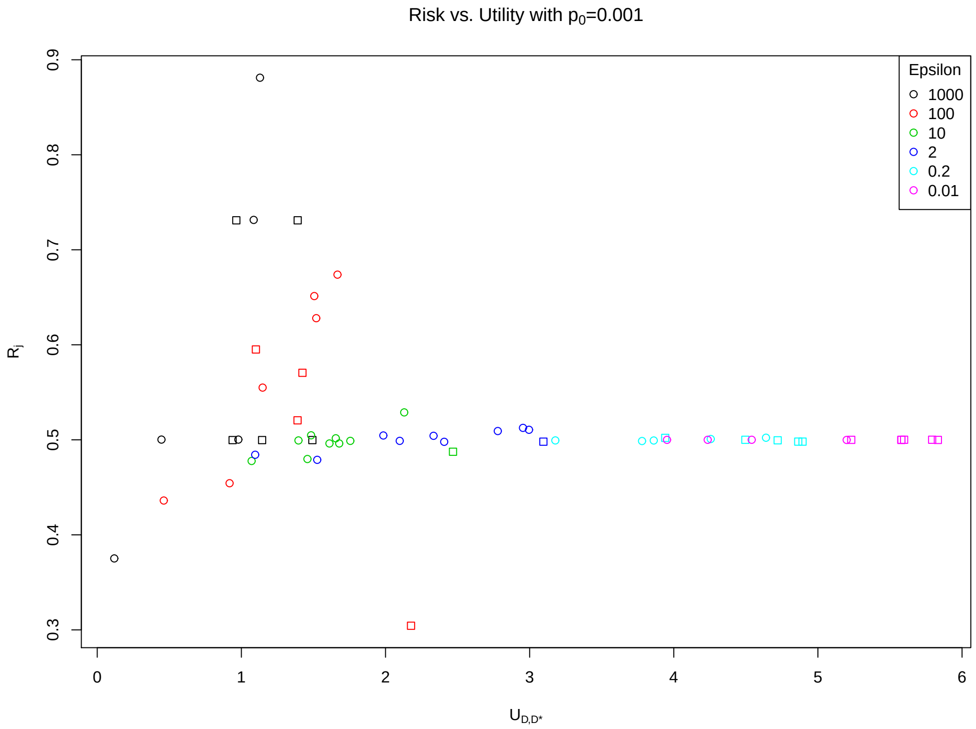

Figure 2: Absolute probabilistic disclosure risks R_j versus data utility $U_{(D,D^*)}$ when $p_0 = .001$. Each point represents a single (D, D^*) pair. Ten data sets are simulated for each ϵ level. The squares represent replicates of D having all values equal to zero, so that we instead compute risks for one record with $Y_j = 0$. Since $p(Y_j = 1 \mid D_{-j}, S) = .5$, $R_j \approx .5$ indicates the intruder learns very little from the data.

3.3 Evaluations of Relative Probabilistic Disclosure Risk

Absolute disclosure risk measures like (2) are sensitive to the specification of $p(Y_j|A,S)$. For example, when the intruder sets $p(Y_j = 1|D_{-j},S) = .9999$, her posterior probability will be relatively large for most D and D^*. Likewise, when the intruder sets $p(Y_j = 1|D_{-j},S) = .0001$, her posterior probability will be relatively small for most D and D^*. We note that ϵ-differential privacy does not suffer from this sensitivity, since it computes a ratio of probabilities over all possible sets of intruder prior knowledge.

We believe that it is prudent to evaluate confidentiality protection schemes accounting for sensitivity to intruders' prior information. We propose to do so with a relative probabilistic disclosure risk, defined generically for a given prior distribution $p(Y_j \mid A,S)$ as

$$RR_j = p(Y_j \mid D^*, A, S)/p(Y_j \mid A, S). \tag{8}$$

This represents the amount intruders can learn from D^* beyond their prior beliefs as encoded in $p(Y_j \mid A,S)$. We note that, for any $p(Y_j \mid A,S)$, RR_j is bounded above when Y_j is discrete since $p(Y_j \mid D^*, A, S)$ reaches a maximum of one. In our simulations, RR_j is simply R_j from (5) divided by the prior probability.

Figures 3 and 4 display plots of the values of RR_j and R_j over a dense grid of possible values of prior beliefs w for the two simulation scenarios in Section 3.2. When $p_0 = .5$, RR_j never exceeds 1.25 regardless of ϵ and w. For these data, D^* does not offer much additional information about Y_j. This story changes when $p_0 = .001$. Here, for small w and $\epsilon \geq 100$, RR_j can exceed two, indicating that the intruder has doubled their posterior probability that $Y_j = 1$ by using D^*. When w is near one, the magnitude of ϵ has little impact on RR_j. Once again, these results suggest that data disseminators using differential privacy may benefit from considering the population characteristics when selecting ϵ.

4 Alternative Disclosure Risk Assessment Approach

Since R_j and RR_j can depend greatly on the nature of the released and original data for given ϵ, it is reasonable to consider alternatives to differential privacy that are specific to the data at hand. However, we also would like to incorporate the idea of (nearly) arbitrary intruder knowledge from differential privacy in these measures. In this section, we outline a statistical risk assessment approach that aims to do so.

Arguably, neither (2) nor (8) are adequate alone, in that data disseminators might care about both the absolute disclosure risk and amount of information gained from D^*. For example, the data disseminator might decide that is acceptable for $RR_j \leq 10$ as long $R_j \leq .10$; that is, it is willing to let the intruder learn up to ten times as much information over her original beliefs (but not more) by using D^* as long as the absolute probabilistic risk does not exceed 10%. Similarly, the data disseminator might decide that it is acceptable for $R_j \geq .90$ as long as $RR_j \leq 1.05$; that is, the data disseminator does not mind if intruders have high absolute risks of disclosure if D^* does not drive where those risks come from, i.e., the intruder's prior distribution explains the high R_j.

We propose that data disseminators consider RR_j and R_j jointly as follows. First, for any Y_j, the data disseminator determines acceptable regions of the $RR_J \times R_j$ space. For example, the data disseminator might accept high RR_j for very low values of R_j, but not accept high values of RR_j for values of $R_j \geq .2$. Second, for the candidate D^*, the data disseminator evaluates (RR_j, R_j) for a range of intruder prior distributions, for example making

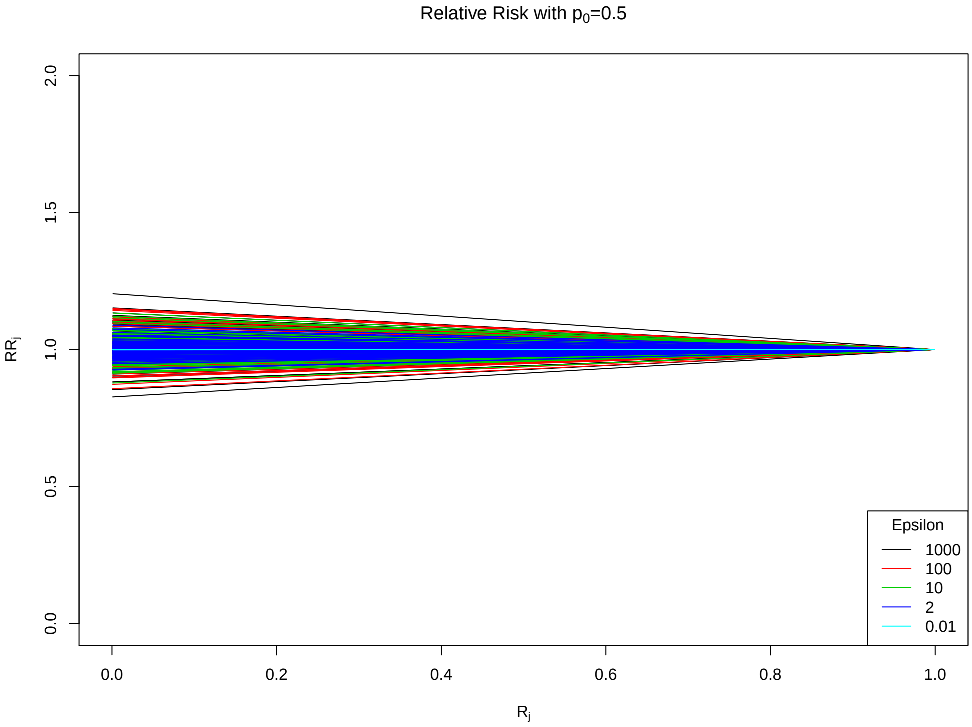

Figure 3: Plots of relative (RR_j) versus absolute (R_j) probabilistic disclosure risks when $p = .5$. Each line represents a single (D, D^*) pair, with RR_j and R_j computed by letting $P(Y_j = 1 \mid D_{-j}, S)$ range over 1000 equally spaced values in $(0, 1)$.

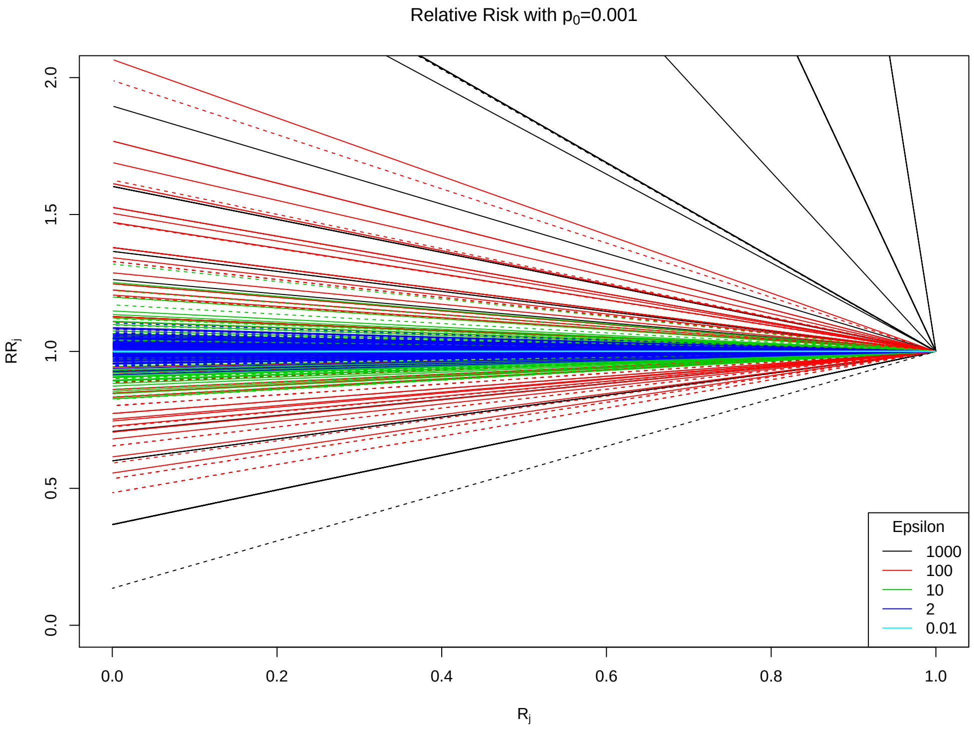

Figure 4: Plots of relative (RR_j) versus absolute (R_j) probabilistic disclosure risks when $p = .001$. Each line represents a single (D, D^*) pair, with RR_j and R_j computed by letting $P(Y_j = 1 \mid D_{-j}, S)$ range over 1000 equally spaced values in $(0, 1)$. Dotted lines represent replicates of D having all values equal to zero, so that we instead compute risks for one record with $Y_j = 0$.

graphical displays akin to those in Figure 3 and Figure 4. Third, the data disseminator determines which, if any, values of w lead to unacceptable joint risks. If all prior distributions result in acceptable (RR_j, R_j), the candidate D^* can be considered safe with respect to the confidentiality of Y_j. If some prior distributions result in unacceptable (RR_j, R_j), the data disseminator either can (i) construct and evaluate other candidates D^*, for example reducing ϵ in our context, or (ii) decide to live with the risk, for example if the data disseminator deems the value of w unlikely to justify in practice. When multiple D^* satisfy the joint risk criteria, the data disseminator can select the generation procedure with the greatest utility, e.g., the largest ϵ in our setting. We note that this approach does not satisfy differential privacy; rather, it is an alternative that seeks to combine the assumption of strong prior knowledge by the intruder with the interpretability and data-specific relevance of statistical risk measures.

In theory, this approach could be scaled up to more realistic settings, although doing so would involve computational challenges. For typically multivariate Y, the data disseminator must compute $p(D^*|D_{-j}, Y_j, S)$ for the true Y_j, which is computationally intensive in high dimensions and with complex D^* generators. In addition, to compute the normalizing constant in (2), the data disseminator must compute this likelihood for other possible values of Y_j. The data disseminator also must specify prior distributions on the potentially very large space of possible values of Y_j. Finally, the data disseminator would need to repeat all of this for all records j considered at risk and combine results in an interpretable manner, e.g., the fraction of times that records are not in the acceptable regions.

While these are serious challenges, there may be reasonable, approximate solutions to some of them. For example, when specifying prior distributions for Y_j, one approach is to put a point mass on the actual Y_j and distribute the remaining probability uniformly over the other possible values of Y_j (or even just over D_{-j} as an approximation to all other possible values of Y_j and hence to the normalizing constant). Or, it may be possible to determine some "worst case"alternatives to the actual value of Y_j, e.g., a record with completely different components of Y_j than the truth, and perform the evaluation using only those worst case alternatives. With regard to evaluation for many j, the computations are trivially parallelizable, so that this should not be a barrier even for large files.

This alternative approach possesses some of the advantages of differential privacy in that (i) it protects against an intruder with access to all but one record, and (ii) it protects against an intruder with arbitrary prior information about the specific variables in Y_j. The approach has some of the advantages of probabilistic risk measures in that (i) it is specific to D^* and D and is not estimated over datasets that were not observed, which arguably are irrelevant, (ii) it results in record-level risk measures, and (iii) it can be adapted to incorporate identification disclosure risks and random sampling. However, the data disseminator must specify intruder knowledge about which variables comprise Y_j, which could range from knowing everything to knowing only certain key variables. Additionally, we suspect that the data disseminator would have to keep secret the acceptable regions of Y_j. Releasing them along with ϵ could leak information about Y_j; for example, intruders could try different values of Y_j until finding one that results in satisfactory regions. We speculate that this would be a challenging attack to implement in practice for data generated from posterior predictive distributions because of the randomness inherent in D^*. Nonetheless, this is a weakness from the perspective of differential privacy.

5 Discussion

Much of the literature on differential privacy focuses on interactive output perturbation rather than noninteractive microdata release. We chose to examine the connections between differential privacy and statistical disclosure risk measures with microdata generation rather output perturbation for two reasons. First, the most prominent application of differential privacy (more precisely, $\epsilon - \delta$ differential privacy) to date in a real-world context is the creation of OnTheMap (Machanavajjhala *et al.*, 2008). This uses a multinomial distribution with prior distributions altered to have differential privacy, which is an extension of the binary synthesizer that we use. We are not aware of any output perturbation systems being heavily used for the release of public use data. Second, we are skeptical that interactive output perturbation systems, at least with current differential privacy technology, can be used for large-scale, public use data dissemination. They have to enforce a finite number of queries, which seems to us a serious drawback for public use datasets in which thousands of individuals may each make hundreds of queries. Differentially private microdata, on the other hand, do not suffer from this limitation.

Reviewers of this manuscript questioned whether or not it is possible to generate synthetic data that are differentially private and analytically useful in genuine settings. Although OnTheMap offers an existence proof, we agree with the reviewers that it is extremely challenging to do so, especially for data with many mixed continuous and categorical variables, data in which estimands based on small samples are of key interest, and data with complex dependencies. There is a long way to go before differentially private synthetic data generation becomes feasible for highly complex datasets. We note, however, that one can generate synthetic microdata (or other microdata) without formally incorporating differential privacy by sampling from predictive distributions, as has been done, for example, in the public use synthetic Longitudinal Business Database (Kinney *et al.*, 2011). Indeed, finding a risk measure for such products motivated the methods in Section 4.

In the simulations, we used the simple setting of binary data primarily to minimize the complexities of data generation. The simulations indicate that the level of ϵ corresponding to particular levels of statistical disclosure risk depends on the properties of the observed data. We are confident that this conclusion will hold in larger and more complex datasets. However, we are not certain what values of ϵ will generate reasonable statistical disclosure risk levels in particular complex datasets: it could be large or small values. Additionally, the risks depend on the nature of the data synthesizer. For example, synthesis models that include many high order interactions be very sensitive to the original data, whereas synthesis models without those interactions may not be.

We believe that ideas from ϵ-differential privacy and statistical disclosure risk measures can be usefully combined in disclosure risk assessment. The proposed alternative risk assessment paradigm in Section 4 represents a step in that direction, and we hope that it generates additional research on integrating the cryptographic and statistical perspectives on risk assessment.

Acknowledgments

This research was supported by a grant from the US National Science Foundation (CNS 1012141).

References

Abowd, J. and Vilhuber, L. (2008). How protective are synthetic data? In J. Domingo-Ferrer and Y. Saygun, eds., *Privacy in Statistical Databases*, 239–246. New York: Springer-Verlag.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). Privacy, accuracy, and consistency too: a holistic solution to contingency table release. *Proceedings of the 27th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems*.

Blum, A., Ligett, K., and Roth, A. (2008). A learning theory approach to non-interactive database privacy. *Proceedings of the 40th ACM SIGACT Symposium on Thoery of Computing*.

Caiola, G. and Reiter, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy* **3**, 27–42.

Charest, A. S. (2010). How can we analyze differentially private synthetic datasets. *Journal of Privacy and Confidentiality* **2:2**, Article 3.

Drechsler, J., Bender, S., and Rässler, S. (2008). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. *Transactions on Data Privacy* **1**, 105–130.

Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In J. Domingo-Ferrer and Y. Saygin, eds., *Privacy in Statistical Databases (LNCS 5262)*, 227–238. New York: Springer-Verlag.

Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association* **81**, 10–28.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics* **7**, 207–217.

Dwork, C. (2006). Differential privacy. In *33rd International Colloquium on Automata, Languages, and Programming, part II*, 1–12. Berlin: Springer.

Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics* **13**, 75–89.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review* **79**, 363–384.

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

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. (2005a). Estimating identification risks in microdata. *Journal of the American Statistical Association* **100**, 1103–1113.

Reiter, J. P. (2005b). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* **168**, 185–205.

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality* **1**, 99–110.

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

Sarathy, R. and Muralidhar, K. (2011). Evaluating Laplace noise addition to satisfy differential privacy for numeric data. *Transactions on Data Privacy* **4**, 1–17.

Skinner, C. J. and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *Journal of the American Statistical Association* **103**, 989–1001.

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association* **105**, 375–389.

Appendix A: Proof that Simulation Scheme Induces ϵ-Differential Privacy

Theorem 1. Let $D = \{x_1, ..., x_n\}$ be a vector of n binary values, and let X_D be the sum of elements in D. Let the randomizer function $f : D \to f(D)$ be such that $f(D) = D^* = \{x_1^*, ..., x_{n_s}^*\}$ is a vector of n_s synthetic binary values (not necessarily the same as n) generated as $x_i^* \overset{iid}{\sim} Bern\left(\frac{X_D+\alpha_\epsilon}{n+\alpha_\epsilon+\beta_\epsilon}\right)$, where $\alpha_\epsilon, \beta_\epsilon \in \mathbb{R}^+$. Let X_{D^*} be the sum of elements in the generated D^*.

For any $\epsilon > 0$, if $\alpha_\epsilon = \beta_\epsilon = \frac{1}{e^{\epsilon/n_s}-1}$, then releasing $f(D) = D^*$ maintains ϵ-level differential privacy. In other words, for any possible D_1, D_2 that differ by at most one element (meaning $|X_{D_1} - X_{D_2}| \leq 1$ in this case) and for any X_{D^*},

$$log\left(\frac{P(X_{D^*}|X_{D_1}, \alpha_\epsilon, \beta_\epsilon)}{P(X_{D^*}|X_{D_2}, \alpha_\epsilon, \beta_\epsilon)}\right) \leq \epsilon$$

Proof: We prove this in two steps:

1. Showing for a fixed α and β that we have

$$\epsilon = \max_{D^*, D_1, D_2}\left(log\left(\frac{P(X_{D^*}|X_{D_1}, \alpha, \beta)}{P(X_{D^*}|X_{D_2}, \alpha, \beta)}\right)\right) = n_s * log\left(\frac{1 + min(\alpha,\beta)}{min(\alpha,\beta)}\right). \quad (9)$$

2. Solving for the value of $\alpha = \beta$ that corresponds to the value of ϵ.

Step 1. Let $\alpha, \beta > 0$. The only way a single element of D can be different is by switching a 0 to 1 (or vice versa). Hence, all possible D_1 and D_2 that differ by one element are considered by examining $X_{D_1} = X+1$, $X_{D_2} = X$ and $X_{D_1} = X$, $X_{D_2} = X+1$ for all $X \in [0 : n-1]$. Let $e_1 = P(X_{D^*}|X+1, \alpha, \beta)$, and let $e_0 = P(X_{D^*}|X, \alpha, \beta)$. The first expression in (9) becomes

$$max\left(\underset{X_{D^*}\in[0:n_s],X\in[0:n-1]}{max}\left(log\left(\frac{e_1}{e_0}\right)\right), \underset{X_{D^*}\in[0:n_s],X\in[0:n-1]}{max}\left(log\left(\frac{e_0}{e_1}\right)\right)\right),$$

which is the same as

$$max\left(\underset{X_{D^*}\in[0:n_s],X\in[0:n-1]}{max}\left(log\left(\frac{e_1}{e_0}\right)\right), \frac{1}{\underset{X_{D^*}\in[0:n_s],X\in[0:n-1]}{min}\left(log\left(\frac{e_1}{e_0}\right)\right)}\right). \tag{10}$$

When finding the values of X and X_{D^*} that maximize or minimize $log\left(\frac{P(X_{D^*}|X+1,\alpha,\beta)}{P(X_{D^*}|X,\alpha,\beta)}\right)$, we can consider the quantity $\frac{P(X_{D^*}|X+1,\alpha,\beta)}{P(X_{D^*}|X,\alpha,\beta)}$ for simplicity, as the log is a monotonically increasing function. We have

$$\begin{aligned}\frac{P(X_{D^*}|X+1,\alpha,\beta)}{P(X_{D^*}|X,\alpha,\beta)} &= \frac{\binom{n}{X_{D^*}}(\frac{X+1+\alpha}{n+\alpha+\beta})^{X_{D^*}}(\frac{n-(X+1)+\beta}{n+\alpha+\beta})^{n_s-X_{D^*}}}{\binom{n}{X_{D^*}}(\frac{X+\alpha}{n+\alpha+\beta})^{X_{D^*}}(\frac{n-X+\beta}{n+\alpha+\beta})^{n_s-X_{D^*}}} \\ &= \left(\frac{X+1+\alpha}{X+\alpha}\right)^{X_{D^*}}\left(\frac{n-X-1+\beta}{n-X+\beta}\right)^{n_s-X_{D^*}}.\end{aligned} \tag{11}$$

Note that $\left(\frac{X+1+\alpha}{X+\alpha}\right) > 1 > \left(\frac{n-X-1+\beta}{n-X+\beta}\right)$. Thus, if we want to maximize (11), we set $X_{D^*} = n_s$ and try to maximize $\left(\frac{X+1+\alpha}{X+\alpha}\right)$, which is done with $X = 0$. Conversely, if we want to minimize (11), we set $X_{D^*} = 0$ and try to minimize $\left(\frac{n-X-1+\beta}{n-X+\beta}\right)$, which is done with $X = n-1$. Substituting these values into (10), we have

$$\begin{aligned}\epsilon &= log\left(max\left(\left(\frac{1+\alpha}{\alpha}\right)^{n_s}, 1/\left(\frac{n-(n-1)-1+\beta}{n-(n-1)+\beta}\right)^{n_s}\right)\right) \\ &= log\left(max\left(\left(\frac{1+\alpha}{\alpha}\right)^{n_s}, \left(\frac{1+\beta}{\beta}\right)^{n_s}\right)\right) \\ &= n_s * log\left(\frac{1+min(\alpha,\beta)}{min(\alpha,\beta)}\right)\end{aligned}$$

This proves that for a given α and β, the DP level will be $\epsilon = n_s * log\left(\frac{1+min(\alpha,\beta)}{min(\alpha,\beta)}\right)$.

Step 2. Set $\alpha = \beta$, and find the α that satisfies the condition for the given ϵ. We have

$$\begin{aligned}\epsilon &= n_s * log(\frac{1+\alpha}{\alpha}) \\ e^{\epsilon/n_s} &= \frac{1+\alpha}{\alpha} \\ \alpha * e^{\epsilon/n_s} - \alpha &= 1 \\ \alpha &= \frac{1}{e^{\epsilon/n_s}-1}.\end{aligned}$$

Thus, when $\alpha_\epsilon = \beta_\epsilon = \frac{1}{e^{\epsilon/n_s}-1}$, the randomizer function offers ϵ-level differential privacy.

Appendix B: Details of Methods for Utility Analysis

In any utility analysis, we require the posterior distribution of p given D^* and S, where S is the information about the synthesis model (including α_ϵ, β_ϵ, and the generating distribution for D^*). Since the sum of the observations in D^* is a sufficient statistic for p, it is equivalent to compute $p(p|X^*, S)$. This posterior distribution is not analytically tractable, so we sample from it using MCMC. We make the sampling process easier by drawing samples from the augmented posterior $p(p, X|X^*, S)$. We write the MCMC for the generic prior distribution, $p \sim Beta(\alpha, \beta)$; we set $\alpha = \beta = 1$ in the simulation studies. The MCMC proceeds as follows.

1. Pick initial values for p and X, which we label $p^{(0)}$ and $X^{(0)}$. We choose $p^{(0)}$ from a $unif(0, 1)$ and $X^{(0)}$ from a uniform distribution on the integers from 0 to n.

2. For $t = 1, \dots, T$, iteratively draw from the two full conditional distributions as follows.

   2.1. *Draw* $X^{(t)} \sim p(X|X^*, p^{(t-1)}, S)$. This is not a closed form distribution, so we use a Metropolis-Hastings step. The proposal distribution, Q, is a one step random walk, i.e., the probability of proposing X' when the chain is currently at X is

$$Q(X \to X') = \begin{cases} 1/2 & \text{if } X' = X + 1 \\ 1/2 & \text{if } X' = X - 1 \\ 0 & \text{otherwise} \end{cases}$$

   for $2 \le X \le n - 1$, and $Q(X = 0 \to X' = 1) = Q(X = n \to X' = n - 1) = 1$. The ratio for accepting X' is thus

$$\begin{aligned} r &= min\left(1, \frac{p(X'|X^*, p^{(t-1)}, S)Q(X' \to X)}{p(X|X^*, p^{(t-1)}, S)Q(X \to X')}\right) \\ &= min\left(1, \frac{p(X^*|X', S)p(X'|p^{(t-1)})Q(X' \to X)}{p(X^*|X, S)p(X|p^{(t-1)})Q(X \to X')}\right) \\ &= min\left(1, \frac{Bin(X^*, n, \frac{X'+\alpha_\epsilon}{n+\alpha_\epsilon+\beta_\epsilon}) * Bin(X', n, p^{(t-1)}) * Q(X' \to X)}{Bin(X^*, n, \frac{X+\alpha_\epsilon}{n+\alpha_\epsilon+\beta_\epsilon}) * Bin(X, n, p^{(t-1)}) * Q(X \to X')}\right). \end{aligned}$$

   We accept or reject X' according to $p(X^{(t)} = X') = r$ and $p(X^{(t)} = X^{(t-1)}) = 1 - r$.

   2.2. *Draw* $p^{(t)} \sim p(p|X^{(t)}, X^*, S)$. With X fixed, from standard Bayesian theory we have $p(p|X^{(t)}) = Beta(X^{(t)} + \alpha, n - X^{(t)} + \beta)$. Hence, $p^{(t)}$ can be drawn directly from this Beta distribution.

These are all relatively fast computations, so that T can be made as large as one hundred thousand without much computation time. We used a burn in period of $B = 1000$, resulting in the samples $p^{(B+1)}, \dots, p^{(T)}$ from $p(p|X^*, S)$.

After obtaining samples from the posterior distribution of p, we seek to compute the numerator and denominator of the utility measure, $U_{(D,D^*)}$. Recall that p_0 is a constant. To compute the numerator, $E((p_0 - p)^2|D^*)$, we substitute the samples $p^{(B+1)}, \dots, p^{(T)}$ into

$\sum_{t=B+1}^{T}(p^{(t)} - p_0)^2/(T - (B + 1))$, where $T = 100,000$ and $B = 1,000$. We compute the denominator, $E((p_0 - p)^2|D)$, via analytical results. We have

$$E((p_0 - p)^2|D) = E(p^2|D) - 2p_0E(p|D) + p_0^2. \tag{12}$$

From standard Bayesian theory, with a uniform prior distribution on p, the posterior distribution of p given D is Beta$(1 + X, n - X + 1)$. Since the expected value and variance of a Beta(a, b) distribution are $\frac{a}{a+b}$ and $\frac{ab}{(a+b)^2(a+b+1)}$, respectively, we can substitute $a = 1 + x$ and $b = n - X + 1$ and solve for all the moments in (12).