How Protective Are Synthetic Data?

John M. Abowd[1] and Lars Vilhuber[1]

School of Industrial and Labor Relations, Cornell University

Abstract. This short paper provides a synthesis of the statistical disclosure limitation and computer science data privacy approaches to measuring the confidentiality protections provided by fully synthetic data. Since all elements of the data records in the release file derived from fully synthetic data are sampled from an appropriate probability distribution, they do not represent "real data," but there is still a disclosure risk. In SDL this risk is summarized by the inferential disclosure probability. In privacy-protected database queries, this risk is measured by the differential privacy ratio. The two are closely related. This result (not new) is demonstrated and examples are provided from recent work.

1 Introduction

When Rubin (1993) introduced the idea of fully synthetic data , there was considerable appeal to releasing data that represented "no actual individual's" responses, and skepticism regarding its feasibility. Subsequent research has adequately demonstrated the feasibility. However, the basic question "How much protection does the synthetic data methodology provide?" remained largely unanswered. The reason is basic: statistical disclosure limitation (SDL) did not provide an adequate framework to answer the question. In the intervening 15 years, a well-developed methodology emerged in the computer science (CS) literature on privacy in databases that allows a synthesis of the techniques used in disclosure limitation and privacy-preserving data mining. The key to this synthesis is the recognition that the privacy measures proposed by the computer scientists and the statistical disclosure limitation methods share a common fundamental– the conditional distribution of the release data, given the underlying confidential data. This short paper provides a roadmap and some examples for moving between the SDL and CS concepts that relate to measuring the protection afforded by synthetic data and the resulting analytical validity of the release data.

2 Definitions

Let X represent a confidential database organized as n rows and k columns of a database table. For clarity in this exposition, assume that only discrete variables may be released and that all variables have been coded as binary outcomes (*e.g.*, yes-no answers). Although one of the great conceptual advantages of fully synthetic data is the possibility of combining continuous and discrete variables,

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 239–246, 2008.

there is no loss of generality in the assumption that the release data consist of contingency tables because all interactions up to k-way are allowed and there are no restrictions on the underlying probabilities. As we will see below, there are practical restrictions on the direct application of these techniques to databases where k is large. We are not going to discuss sampling as a disclosure limitation technique; consequently, we will assume that n is the population and $n_i = 1$ is a population unique. That is, there is one, and only one row of X in which i^{th} column has a 1.

Let $\boldsymbol{\pi}$ be the $(k \times 1)$ vector of probabilities associated with the complete table, where all elements of $\boldsymbol{\pi}$ are strictly positive. Assume that the contingency table is summarized by a vector of counts $\boldsymbol{n}$ that is also $(k \times 1)$ with $n = \sum_{i=1}^{k} n_i$. Without loss of generality, assume that the confidential data are distributed Multinomial, $\boldsymbol{n} \sim \mathrm{M}(n, \boldsymbol{\pi})$. Summarize all prior information about the parameters by assuming that they are drawn from a Dirichlet distribution, $\boldsymbol{\pi} \sim \mathrm{D}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is the $(k \times 1)$ vector of prior sample sizes with $\alpha_0 = \sum_{i=1}^{k} \alpha_i$. Then, the posterior predictive distribution of the confidential data can be constructed by noting that $\boldsymbol{\pi} \sim \mathrm{D}(\boldsymbol{\alpha} + \boldsymbol{n})$ *a posteriori*.

Let $\tilde{X}$ denote a single synthetic data set based on X. Suppose that $\tilde{X}$ is $(m \times k)$. The synthetic data can be constructed by first sampling $\tilde{\boldsymbol{\pi}} \sim \mathrm{D}(\boldsymbol{\alpha} + \boldsymbol{n})$, then constructing the rows of $\tilde{X}$ from counts sampled from $\mathrm{M}(m, \tilde{\boldsymbol{\pi}})$. Because of the way $\tilde{X}$ is constructed, we can represent the conditional distribution of $\tilde{X}$ given X using

$$\Pr[\boldsymbol{m}|\boldsymbol{n}, MD] = \mathrm{E}_{\boldsymbol{\pi}|\boldsymbol{n}}[\mathrm{M}(m, \boldsymbol{\pi})|MD] \tag{1}$$

where we have noted explicitly that the conditional distribution depends upon the Multinomial-Dirichlet (MD).

The argument leading up to the construction of $\Pr[\boldsymbol{m}|\boldsymbol{n}, MD]$ above is a complete Bayesian analysis, and equation (1) defines the posterior predictive distribution of $\tilde{X}$ given X. But the Bayesian analysis is not essential to the synthetic data construction. Any transition function $\Pr[\boldsymbol{m}|\boldsymbol{n}]$ that defines a proper conditional distribution for the synthetic counts given the confidential counts can be used to synthesize data. Dwork *et al.* (2006) define a synthesizer for the same confidential database problem by sampling k i.i.d. random variables from the Laplace (double exponential) distribution $\mathrm{Lap}(0, 2/\epsilon)$, where the reason for defining the scale parameter in the form shown will be made clear below. Let $\boldsymbol{y}$ be the $(k \times 1)$ vector of Laplacian random variables. Define the synthetic counts as $\boldsymbol{m} = \boldsymbol{n} + \boldsymbol{y}$. Using the properties of the Laplace distribution, they construct an alternative conditional distribution

$$\Pr[\boldsymbol{m}|\boldsymbol{n}, Lap] = \Pr[\boldsymbol{n} + \boldsymbol{y}|\boldsymbol{n}, \epsilon]. \tag{2}$$

The above discussion has been in terms of conditional distributions. A generic random sanitizer is defined as any function $\tilde{X} \leftarrow \mathrm{San}(X, Y)$ that maps the confidential data X and random noise Y of specified dimensionality into a sanitized

copy of the database, denoted $\tilde{X}$ here to emphasize its relation to synthetic data. Because of the way we constructed $\boldsymbol{n}$ from X, there is a completely equivalent sanitizer $\boldsymbol{m} \leftarrow \text{San}(\boldsymbol{n}, \boldsymbol{y})$. Hence, any sanitizer can be used to construct a conditional distribution $\Pr[\boldsymbol{m}|\boldsymbol{n}, San]$. Thus, a discussion of sanitizers is equivalent to a discussion of the conditional distribution constructed from those sanitizers, and in the remainder of this paper, we will focus on conditional distributions, without loss of generality.

3 Statistical Disclosure Limitation and Differential Privacy

Consider a generic conditional distribution $\Pr[\boldsymbol{m}|\boldsymbol{n}]$, and represent the conditional probabilities in a matrix Υ $(k \times k)$. SDL methods focus on the rows of Υ. For example, if $\Upsilon = I$, then the release data are identical to the confidential data. If

$$\max(\text{diag}(\Upsilon)) < 1 - \delta;$$

then, the release data differ from confidential data in every dimension by at least δ. That is, for all $i = 1, \ldots k$

$$\Pr[m_i \neq n_i|\boldsymbol{n}, San] > \delta$$

and the SDL is defined to have infused at least $\delta-$percent uncertainty into every tabulation. Acceptable levels of δ are usually an inverse function of n_i. Furthermore, the actual values of δ are usually kept secret.

By contrast, the computer science data privacy literature concerns itself with the columns of Υ. To understand this formally, consider two copies of X, say $X^{(1)}$ and $X^{(2)}$ that differ in a single row such that $\left|\boldsymbol{n}^{(1)} - \boldsymbol{n}^{(2)}\right| = 2$. While this condition looks obscure, it amounts to assuming that the two copies of the database differ on a single attribute of a single row; hence, some n_i changes from 0 to 1 while exactly one other n_j changes from 1 to 0. Dwork *et al.* (2006) define $\epsilon-$differential privacy as the requirement that

$$\max \left| \ln \left(\frac{\Pr\left[\boldsymbol{m}|\boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{m}|\boldsymbol{n}^{(2)}\right]} \right) \right| \leq \varepsilon \tag{3}$$

where the max is taken over $\forall \boldsymbol{n}^{(1)}, \boldsymbol{n}^{(2)}$ where $\left|\boldsymbol{n}^{(1)} - \boldsymbol{n}^{(2)}\right| = 2$ and all columns of Υ respecting the convention that the larger element is placed in the numerator.[1] Thus, the computation of the ratios of elements of each column of Υ considers only those combinations for the numerator and denominator that can be reached by change of a single row of X. As an enhancement, Machanavajjhala *et al.* (2008) define $(\epsilon, \delta)-$probabilistic differential privacy as the requirement that equation (3) hold with probability $1 - \delta$ for $\forall \boldsymbol{n}^{(1)}, \boldsymbol{n}^{(2)}$ where $\left|\boldsymbol{n}^{(1)} - \boldsymbol{n}^{(2)}\right| = 2$, where the

[1] Dwork *et al.* (2006) actually call this $\epsilon-$indistinguishability. Dwork (2006) standardizes the terminology to $\epsilon-$differential privacy.

probabilities are calculated with respect to the joint distribution of $(\boldsymbol{m}, \boldsymbol{n})$, given $\boldsymbol{\alpha}$. They interpret probabilistic differential privacy as $\epsilon-$differential privacy that fails with probability δ, a rare event.

Conventional SDL methods and differential privacy definitions are related by the concept of an inferential disclosure. An inferential disclosure occurs when the attacker can infer the value of a variable for a row in the confidential data by comparing the release data to the information available without the release data (the attacker's information set, or prior). The attacker's prior knowledge is summarized by the ratio

$$\frac{\Pr\left[\boldsymbol{n} = \boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{n} = \boldsymbol{n}^{(2)}\right]}$$

which measures the extent to which the attacker can ascertain the difference between $\boldsymbol{n}^{(1)}$ and $\boldsymbol{n}^{(2)}$ without using the release data. The attacker's gain in information from having access to the synthetic release data $\boldsymbol{m} = \tilde{\boldsymbol{m}}$ is given by the posterior odds ratio

$$\frac{\frac{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(1)}|\tilde{\boldsymbol{m}}\right]}{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(2)}|\tilde{\boldsymbol{m}}\right]}}{\frac{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(2)}\right]}}. \tag{4}$$

If the posterior odds ratio is large, then the release data contain a great deal of information about the row associated with the change from $\boldsymbol{n}^{(1)}$ and $\boldsymbol{n}^{(2)}$. At the limit, if this ratio is infinite, an inferential disclosure is certain. But it turns out that

$$\frac{\frac{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(1)}|\tilde{\boldsymbol{m}}\right]}{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(2)}|\tilde{\boldsymbol{m}}\right]}}{\frac{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(2)}\right]}} = \frac{\Pr\left[\boldsymbol{m} = \tilde{\boldsymbol{m}}|\boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{m} = \tilde{\boldsymbol{m}}|\boldsymbol{n}^{(2)}\right]}$$

Hence, $\epsilon-$differential privacy limits the maximum gain in information (posterior odds) for an attacker who knows all properties of the disclosure limitation procedure $(\Pr[\boldsymbol{m}|\boldsymbol{n}])$, and all rows of X save one, to

$$\max\left[\frac{\Pr\left[\boldsymbol{m} = \tilde{\boldsymbol{m}}|\boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{m} = \tilde{\boldsymbol{m}}|\boldsymbol{n}^{(2)}\right]}\right]$$

where the max is taken over $\forall \boldsymbol{n}^{(1)}, \boldsymbol{n}^{(2)}$ where $\left|\boldsymbol{n}^{(1)} - \boldsymbol{n}^{(2)}\right| = 2$ and all columns of Υ. Furthermore, $(\epsilon, \delta) -$ probabilistic differential privacy limits the maximum gain in information for an attacker with this information with probability $1 - \delta$.

We can now answer the question posed in the title. Fully synthetic data, the type we have discussed in this paper, are protective of the confidential data to the extent that they limit inferences of the type defined by equation (4). Hence, synthetic data that display $\epsilon-$differential privacy are guaranteed to be protective against an attacker with full information about the data protection process (knowledge of $\boldsymbol{\alpha}$ and n for $\Pr[\boldsymbol{m}|\boldsymbol{n}, MD]$; knowledge of ϵ but not n

for $\Pr[\boldsymbol{m}|\boldsymbol{n}, Lap]$; knowledge of $\Pr[\boldsymbol{m}|\boldsymbol{n}, San]$, in general) and knowledge of all but one row of X. Similarly, synthetic data that display $(\epsilon, \delta)-$ probabilistic differential privacy are protective against the same attacker with probability $1-\delta$.

Thus, synthetic data that have one of these differential privacy properties protect against an attacker with an enormous information set, certainly containing more information than conventional SDL procedures assume. But, what of synthetic data procedures that do not satisfy differential privacy? A sanitizer that doesn't satisfy either $\epsilon-$differential privacy or $(\epsilon, \delta)-$ probabilistic differential privacy displays infinite differential privacy ($\epsilon \to \infty$) for some kinds of attacks. Virtually every SDL procedure in regular use–suppression, coarsening, swapping, shuffling, sampling, and most noise-infusion techniques–fails to satisfy differential privacy. For this reason, the users of these methods normally safeguard the parameters and conditioning information required to calculate $\Pr[\boldsymbol{m}|\boldsymbol{n}, San]$. However, applying a differential privacy audit to synthesizers and sanitizers in regular use can be very instructive about their strengths and limitations, as we hope the examples below will demonstrate.

4 Applications

4.1 The Multinomial-Dirichlet Synthesizer

Figure 1 displays $\Pr[\boldsymbol{m}|\boldsymbol{n}, MD]$ a Multinomial-Dirichlet synthesizer that has $(2, 0.0006)-$probabilistic differential privacy. The synthesizer displays the entire sample space for $n = 5, k = 2, \alpha_0 = 1.0, \alpha_1 = \alpha_2 = 0.5$. There is no suppression in the output; hence, every combination of actual data (rows) can produce any possible outcome (columns). This synthesizer displays finite differential privacy, as can be seen in Figure 2. It is the eight cells that have values in excess of 2 that cause the failure of strict $\epsilon-$differential privacy, and those cells have a combined probability of 0.0006.

The properties displayed in Figure 1 are generic features of Multinomial-Dirichlet synthesizers that satisfy finite differential privacy. Notice that the cells that have the largest log posterior odds ratios are those in which the synthesizer delivers "unusual" outcomes–outcomes that are far from the sample data. The

n_1	n_2	m_1 0 / m_2 5	1 / 4	2 / 3	3 / 2	4 / 1	5 / 0
0	5	0.647228	0.294194	0.053490	0.004863	0.000221	0.000004
1	4	0.237305	0.395508	0.263672	0.087891	0.014648	0.000977
2	3	0.067544	0.241227	0.344610	0.246150	0.087911	0.012559
3	2	0.012559	0.087911	0.246150	0.344610	0.241227	0.067544
4	1	0.000977	0.014648	0.087891	0.263672	0.395508	0.237305
5	0	0.000004	0.000221	0.004863	0.053490	0.294194	0.647228

Fig. 1. Multinomial-Dirichlet synthesizer with (2,0.0006)-prob. differential privacy

$n_1^{(1)}$	$n_2^{(1)}$	$n_1^{(2)}$	$n_2^{(2)}$	m_1 / m_2	0	1	2	3	4	5
					5	4	3	2	1	0
0	5	1	4		1.003353	0.295930	1.595212	2.894495	4.193778	5.493061
1	4	2	3		1.256572	0.494432	0.267708	1.029848	1.791988	2.554128
2	3	3	2		1.682361	1.009417	0.336472	0.336472	1.009417	1.682361
3	2	4	1		2.554128	1.791988	1.029848	0.267708	0.494432	1.256572
4	1	5	0		5.493061	4.193778	2.894495	1.595212	0.295930	1.003353

Fig. 2. Differential privacy values (log posterior odds ratios) for MD synthesizer

natural tendency is to set the synthesizer so that it suppresses these outcomes, but that technique creates zeros in the rows of the transition matrix and, hence, infinite differential privacy. For these cases, probabilistic differential privacy allows the log posterior odds ratios to be large for exactly the low-probability outcomes of the synthesizer.

4.2 The Laplace Sanitizer

Figure 3 displays $\Pr[\boldsymbol{m}|\boldsymbol{n}, Lap]$ for the same (5×2) data matrix with the parameters of the Laplace distribution chosen to guarantee $2-$differential privacy, as in the example above. In order to make the comparison with the MD synthesizer interesting, We have assumed that the total size of the database, $n = 5$ is known. Hence, the appropriate distribution for the noise is $\text{Lap}\,(0, 2/\epsilon)$ with $\epsilon = 2$ (see Dwork *et al.*, page 8), but there is only one query being protected, not two, since the total number of rows in the database is known. Figure 4 confirms that the transition matrix guarantees $2-$differential privacy.

The Laplace sanitizer displayed in Figure 3 is also typical. It displays larger probabilities for the rare events than the MD synthesizer because it never allows the log odds ratio to exceed 2. But, it is also more peaked around the high-probability transitions, which is a feature of the double exponential noise used in the sanitizer.

n_1	n_2	m_1 / m_2	0	1	2	3	4	5
			5	4	3	2	1	0
0	5		0.816060	0.159046	0.021525	0.002913	0.000394	0.000062
1	4		0.183940	0.632121	0.159046	0.021525	0.002913	0.000456
2	3		0.024894	0.159046	0.632121	0.159046	0.021525	0.003369
3	2		0.003369	0.021525	0.159046	0.632121	0.159046	0.024894
4	1		0.000456	0.002913	0.021525	0.159046	0.632121	0.183940
5	0		0.000062	0.000394	0.002913	0.021525	0.159046	0.816060

Fig. 3. Laplace synthesizer with 2-differential privacy

$n_1^{(1)}$	$n_2^{(1)}$	$n_1^{(2)}$	$n_2^{(2)}$ (m_1 \ m_2)	0	1	2	3	4	5
				5	4	3	2	1	0
0	5	1	4	1.489880	1.379885	2.000000	2.000000	2.000000	2.000000
1	4	2	3	2.000000	1.379885	1.379885	2.000000	2.000000	2.000000
2	3	3	2	2.000000	2.000000	1.379885	1.379885	2.000000	2.000000
3	2	4	1	2.000000	2.000000	2.000000	1.379885	1.379885	2.000000
4	1	5	0	2.000000	2.000000	2.000000	2.000000	1.379885	1.489880

Fig. 4. Differential privacy values (log posterior odds ratios) for Laplace sanitizer

5 Discussion

This short article is just meant to illustrate what is required to answer the question "How protective are synthetic data?" and to provide some generic examples for simple problems. The two articles upon which we have primarily relied contain many more details of both procedures. In particular Machanavajjhala *et al.* (2008) show that the real challenge for the MD synthesizer is to handle problems where the number of columns in the database is huge. Their example, an origin-destination commuting pattern database, has 8.2 million rows. Both the MD synthesizer and the Laplace sanitizer deliver poor analytical validity in this example unless the domain is coarsened. The MD synthesizer gives poor results without coarsening because the minimum prior sample size that must be spread across the 8.2 million possible origins is usually much larger than the number of sample individuals. The Laplace synthesizer also adds noise to each origin and, while the properties of the Laplace noise do not depend upon the number of potential origins (8.2 million), if the release data are provided for each origin, the total amount of noise in the release data is comparable to the M-D synthesizer.

Coarsening the domain can be difficult since all feasible outcomes must have positive transition probabilities for every row of the input database in order to preserve either type of differential privacy. Machanavajjhala *et al.* (2008) address this problem by combining distance-based coarsening with a probabilistic pruning algorithm. When used in combination, the analytical properties of the data can be preserved with a $(4, 0.0001)-$probabilistic differential privacy (Machanavajjhala *et al.*, 2008, page 9).

Dwork *et al.* (2006) consider an equally difficult problem–all possible tables from a census of population. Barak *et al.* (2008) show how to guarantee $\epsilon-$differential privacy by coarsening this problem via a restatement in the Fourier basis, where far fewer free coefficients are required to guarantee privacy.

There are many unsolved problems in the application of formal privacy models and SDL to fully synthetic data. This article illustrates the common ground in the two methodologies and points out ways to implement the procedures in complex data models.

Acknowledgements

We gratefully acknowledge the support of National Science Foundation grants SES-0339191, SES-0427889 and CNS-0627680. We are also indebted to Fredrik Andersson, Cynthia Dwork, Johannes Gehrke, Dan Kifer, Ashwin Machanavajjhala, Kobbi Nissim, Jerry Reiter, and Adam Smith for valuable ongoing discussions of the issues discussed herein.

References

Rubin, D.B.: Discussion of statistical disclosure limitation. Journal of Official Statistics 9, 461–468 (1993)

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K.: Privacy, accuracy, and consistency, too: A holistic solution to contingency table release. In: PODS 2007 (2007)

Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)

Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: Theory meets practice on the map. In: International Conference on Data Engineering, ICDE 2008 (in press, 2008)