Differential Correct Attribution Probability for Synthetic Data: An Exploration

Jennifer Taub(✉), Mark Elliot, Maria Pampaka, and Duncan Smith

The University of Manchester, Manchester M13 9PL, UK
jennifer.taub@postgrad.manchester.ac.uk,
{mark.elliot,maria.pampaka,duncan.smith}@manchester.ac.uk

Abstract. Synthetic data generation has been proposed as a flexible alternative to more traditional statistical disclosure control (SDC) methods for limiting disclosure risk. Synthetic data generation is functionally distinct from standard SDC methods in that it breaks the link between the data subjects and the data such that reidentification is no longer meaningful. Therefore orthodox measures of disclosure risk assessment - which are based on reidentification - are not applicable. Research into developing disclosure assessment measures specifically for synthetic data has been relatively limited. In this paper, we develop a method called Differential Correct Attribution Probability (DCAP). Using DCAP, we explore the effect of multiple imputation on the disclosure risk of synthetic data.

Keywords: Synthetic data · Disclosure risk · CART

1 Introduction

With the increasing centrality of data in our lives, societies and economies, and the drive for greater government transparency and release of open data, there has been a concomitant increase in demand for public release microdata. However, many of the traditional SDC techniques still are subject to disclosure risks. For example, Dinur and Nissim (2003) showed that additive noise (a common SDC method) is not protective against certain kinds of attacks/adversaries and Elliot et al. (2016) demonstrate that standard SDC controlled datasets are vulnerable to reidentification attacks using a jigsaw identification if released as open data.

An alternative to traditional SDC techniques is the use of synthetic data. The idea of synthetic data was first introduced by Rubin (1993), who proposed treating each observed data point as if it were missing and imputing it conditional on the other observed data points to produce a fully synthetic dataset. As an alternative, Little (1993) introduced a method that would replace only the sensitive variables in the observed data, to produce what is referred to as

J. Domingo-Ferrer and F. Montes (Eds.): PSD 2018, LNCS 11126, pp. 122–137, 2018.
https://doi.org/10.1007/978-3-319-99771-1_9

partially synthetic data. Since fully synthetic data does not contain any original data, the disclosure of information from the synthetic data is less likely to occur. Likewise for partially synthetic data, the sensitive values are synthesised, and thus the risk of disclosure of sensitive information is lessened compared to the original data.

Rubin's initial proposal for producing synthetic data was based on multiple imputation (MI) techniques using parametric modelling. Rubin originally developed multiple imputation in the 1970s as a solution to deal with missing data by replacing missing values with multiple values, to account for the uncertainty of the imputed values. Alternatively, synthetic data can be created using single imputation (SI) wherein the uncertainty of a missing value is accounted for by a different set of rules for variance estimation (Raab et al. 2016). Recent research has examined non-parametric methods - including machine learning techniques - which are better at capturing non-linear relationships (Drechsler and Reiter 2011). These methods include classification and regression trees (CART) (Reiter 2005), random forests (Caiola and Reiter 2010), bagging (Drechsler and Reiter 2011), support vector machines (Drechsler 2010), and genetic algorithms (Chen et al. 2016). In this paper we will be using synthetic data generated from both traditional parametric modelling and from CART.

Synthetic data generation is still a relatively new method of data protection and there has been a relative dearth of research into assessment of residual disclosure risk. The way that it operates is functionally distinct from standard SDC methods in that it breaks the link between the data subjects and the output data in a way that SDC methods do not. Thus a common sense view might posit that synthetic data are without disclosure risk since the data do not represent real individuals. However, synthetic data may pose a disclosure risk through the attributions or inferences that it allows. Therefore, orthodox measures for disclosure risk which are based on the notion of re-identification are not applicable for fully synthetic data. Research into measuring risk for synthetic data has been limited to a few papers (notably Reiter and Mitra (2009); Reiter et al. (2014)).

In this paper we will be developing and using a measure introduced procedurally by Elliot (2014) to capture the disclosure risk of multiply imputed parametric and CART synthetic data. The remainder of this article is structured as follows: Sect. 2 presents an overview of disclosure risk with a focus on disclosure risk for synthetic data. Section 3 discusses the dataset being used, how the synthetic files are generated, and how we use DCAP. Section 4 presents the results of our analyses and discusses the relationship between disclosure risk and number of imputations, leading to our conclusion in Sect. 5.

2 Disclosure Risk

Data disclosure can occur in different forms. The most important forms are re-identification and attribution. Re-identification occurs when an identity is attached to a data unit, while attribution is when some attribute can be associated with a population unit (eg. Elliot et al. 2016). For example, attribution might occur if some microdata reveal that all men age 65+ in a particular geographical area have prostate cancer. Therefore, a data intruder who lives in that area would learn that their 65+ male neighbour has prostate cancer, which is information that the survey was clearly not meant to disclose. Re-identification and attribution frequently occur together but can occur separately. Some authors also distinguish another type, inference attacks, wherein an attacker can infer information at a high degree of confidence[1]. However, we would argue that the distinction is somewhat arbitrary since all attributions involve some degree of uncertainty. We instead define attributions as inferences that a population unit has a certain characteristic and define a subclass of those attributions as *correct attributions*, wherein an intruder correctly identifies an attribute for a given respondent. It follows that a well formed measure of attribution risk would capture the proportion of attributions that are correct.

With synthetic data, as Drechsler et al. (2008) write, "the intruder faces the problem that he never knows (i) if the imputed values are anywhere near the true values and (ii) if the target record is included in one of the different synthetic samples[2]" (p. 1018). Following this reasoning, it is widely understood that thinking of risk within synthetic data in terms of re-identification, which is how many other SDC methods approach disclosure risk, is not meaningful and therefore we must develop measures of attribution risk (see for example Reiter and Mitra (2009)).

2.1 Disclosure Risk Measures for Synthetic Data

Unfortunately, the bulk of the SDC literature (for example: Winkler 2005; Skinner and Elliot 2002; Elliot et al. 2002; Yancey et al. 2002; Fienberg and Makov 1998; Kim and Winkler 1995, etc.) has focused on re-identification in the form of record linkage, which is not meaningful for fully synthetic data. By focusing their efforts on re-identification, they do not address disclosure risk that occurs in the form of attribution without re-identification, such as that addressed by Smith and Elliot (2008) and Machanavajjhala et al. (2007).

Previous methods for synthetic data risk estimation include Reiter et al.'s (2014) Bayesian estimation approach and Reiter and Mitra's (2009) matching probability of partially synthetic data. Reiter and Mitra compare perceived match risk, expected match risk, and true match risk. Reiter et al. assume that an intruder seeks a Bayesian posterior distribution. In Reiter et al.'s framework, the intruder is assumed to know all of the records but one. This scenario is very unlikely to occur in the real world, and even the authors noted that it is a conservative estimate. Essentially this is an approach that overestimates disclosure risk.

Differential Correct Attribution Probability (DCAP). Elliot (2014) introduced an approach that combines the notion that one should measure attribution risk as the probability of an attribution being correct and that one can

[1] The level of confidence which is regarded as disclosive is a subjective judgement.

[2] A synthetic dataset often contains multiple synthetic samples (m).

then compare that probability to those obtained on the original data and against some baseline. Here we develop this approach more formally and will refer to it as the *Differential Correct Attribution Probability* (DCAP).

The underlying measure of DCAP is the correct attribution probability and this has some similarity to Reiter and Mitra's method. Both methods employ a matching mechanism with the assumption that the intruder knows the true values of a key consisting of non-target variables. However, the framing of the two approaches is different. Given that Reiter and Mitra are matching using partially synthetic data, when a match occurs amongst the statistical uniques[3], the intruder can be certain that the match occurring refers to the same record. However, what is uncertain in the Reiter and Mitra method, is whether the synthetic version of a target record is the same as that of the real target. While in DCAP, since the matching is occurring with fully synthetic data, the matching has considerably less certainty since the keys do not directly map onto one another. In Reiter and Mitra's method they are comparing assumed risk against actual matches. While in DCAP the correct attribution rate for synthetic data is compared to that of the original dataset and a baseline univariate distribution. In the DCAP method one record can have multiple matches even with the original dataset and therefore, it is merely a probability of attributing the correct target variable to the key, since it is not only concerned with uniques and nor does it assume verifiable matches, as in Reiter and Mitra's scenario. In essence Reiter and Mitra's method is best for partially synthetic data, while DCAP is better suited for fully synthetic data.

DCAP works on the assumption that the intruder knows the values of a set of key variables for a given unit and is seeking to learn the specific value of a target variable. Where the target variable is categorical[4] the method works as follows: We define d_o as the original data and K_o and T_o as vectors for the key and target information

$$d_o = \{K_o, T_o\} \tag{1}$$

Likewise, d_s is the synthetic dataset.

$$d_s = \{K_s, T_s\} \tag{2}$$

The Correct Attribution Probability (CAP) for the record indexed j is the empirical probability of its target variables given its key variables,

$$CAP_{o,j} = Pr(T_{o,j}|K_{o,j}) = \frac{\sum_{i=1}^{n}[T_{o,i} = T_{o,j}, K_{o,i} = K_{o,j}]}{\sum_{i=1}^{n}[K_{o,i} = K_{o,j}]} \tag{3}$$

where the square brackets are Iverson brackets and n is the number of records.

[3] A statistically unique record is a record in the dataset, in which no other record in the dataset has that particular combination of characteristics.

[4] Elliot (2014) presents a variant where the target is continuous but we do not consider that here.

The CAP for record j based on a corresponding synthetic dataset d_s is the same empirical, conditional probability but derived from d_s,

$$CAP_{s,j} = Pr(T_{o,j}|K_{o,j})_s = \frac{\sum_{i=1}^{n}[T_{s,i} = T_{o,j}, K_{s,i} = K_{o,j}]}{\sum_{i=1}^{n}[K_{s,i} = K_{o,j}]} \tag{4}$$

For any record in the original dataset for which there is no corresponding record in the synthetic dataset with the same key variable values, the denominator in Eq. 4 will be zero and the CAP is therefore undefined. In Sect. 3.3 we describe two methods for dealing with this.

The baseline CAP for record j is the marginal probability of its target variables estimated from the original dataset,

$$CAP_{b,j} = Pr(T_{o,j}) = \frac{1}{n}\sum_{i=1}^{n}[T_{o,i} = T_{o,j}] \tag{5}$$

In principle, the baseline could be set to any level. However, the choice of the univariate baseline in Eq. 5 as the default for the approach, is based on the pragmatic assumption that the intruder will routinely know the univariate distribution of a target variable for the population. Another way to look at this is that a univariate distribution is often considered releasable into the public domain; univariate distributions are frequently published as summary statistics, and so a synthetic CAP score equal to or lower than the baseline CAP is likely to be considered as a sufficiently low risk in most conceivable situations.[5] On the other hand if the synthetic CAP score is equal to the CAP score for the original data that would imply that the synthetic data is as disclosive as the original data and, therefore, the synthetic data generator is not sufficiently protecting the data. The original data and baseline CAP scores therefore represent effective operating bounds in which the risk associated with synthetic data is likely to fall.

We propose addressing multiple imputations for DCAP by pooling the matches from the multiple synthetic samples:

$$CAP_m = Pr(T_1 + ... + T_m = T_o|K_1 + ... + K_m = K_s) \tag{6}$$

where m is indicative of the imputation. This is preferable to averaging because we believe that an actual intruder would be viewing a synthetic dataset in its entirety. This paper will address the following research questions:

1. How does using Multiple Imputation affect the CAP score?
2. How do statistical uniques affect the CAP score?
3. Do parametrically and CART generated data have similar CAP scores?
4. Is synthetic data differentially confidential?

[5] It is worth noting that if the mean CAP score of the whole synthetic dataset is at the baseline, that effectively means that the target is independent of the key which may be indicative that the data have a utility issue.

3 Empirical Demonstation of the DCAP Approach

3.1 Data Sources

Following Elliot (2014) we used the Living Costs and Food Survey (LCF) Office for National Statistics (2016) as one of our test datasets. The LCF has many characteristics that make it a good candidate as a test dataset. First, it has a small size, which will allow for the dataset to be quickly synthesised. Additionally, the LCF has detailed information, making it vulnerable to disclosure. Furthermore, since the LCF was used by Elliot (2014) this allows us to use his keys and targets. We used the 2014 LCF, which consists of 5133 records. We used the following variables: Government office region (GOR), household size, output area classifier, tenure, economic position of reference person, dwelling type, number of workers, number of cars, and, internet in household (synthesised in that order). We utilised three different versions of the LCF: (1) the original 2014 LCF; (2) a CART generated 2014 LCF $m = 10$, and (3) a parametrically generated 2014 LCF $m = 10$. Our second test dataset is the British Social Attitudes Survey (BSA) NatCen Social Research (2014). The 2014 BSA consists of 2878 records. We used the following variables: GOR, higher education qualification, marital status, age category, gender, social class, and household income. We will also use three different versions of the BSA: (1) the original 2014 BSA; (2) a CART generated 2014 BSA $m = 10$, and (3) a parametrically generated 2014 $m = 10$.

3.2 Creation of Synthetic Data Files

The synthetic data are generated using the r-package, synthpop version 1.3-0 (Nowok et al. 2016). Synthpop was used to generate both the parametric and CART synthetic datasets. The parametric synthetic dataset was generated using logistic regression and polytomous logistic regression, since all variables being used are categorical. The missing data from the original LCF and BSA are left unchanged in the synthesis process.

3.3 Parameters for the Experiments

Our key variables for the LCF are as follows: GOR, Output area classifier, tenure, dwelling type, internet in household, and household size. The target variable is economic position of reference person. The first four variables of the key and the target variable are the same as in Elliot (2014). For the BSA the key variables are: GOR, education qualification, marital status, age, gender, social class and household income. The target variable is banded income. (Different key sizes for the LCF are in Appendix A).

Treatment of Non-matches in the CAP Score. DCAP, like many previous disclosure risks measures, works on the basis of matching on key variables. However, here we are not primarily concerned with the status of those matches but whether they lead to correct or incorrect attributions. The CAP score is the proportion of matches leading to correct attributions out of total matches. However, when measuring the CAP score for synthetic data, not every combination of keys from the original dataset will be present in the synthetic dataset. Elliot presented two different resolutions for this issue: recording the CAP values for the non-matches as zero or treating a non-match on the key as undefined, whereby the record is discounted and does not count towards n. The basis for assigning a zero is that a non-match has a zero probability of yielding a correct attribution. However, the logic behind recording non-matches as undefined is that an actual intruder is more likely to stop their attempt with a non-match. Elliot (2014) found that treating non-matches as undefined leads to higher CAP scores. In this paper we will be exploring CAP scores for both when non-matches are recorded as zero and coded as undefined.

Different Intruder Scenarios. When originally proposed, the CAP score was intended to be averaged across the whole dataset, however there is nothing intrinsic to DCAP that requires the use of the entire dataset. DCAP can be used for a variety of different intruder scenarios. However, in all scenarios, it is assumed that the intruder knows the information from the key for the original dataset and that the target variable is unknown to the intruder.

We will be exploring DCAP in three different scenarios, one where the entire dataset is in use, a second where the intruder is only interested in respondents who are statistically unique for the key (this would make it equivalent to the Reiter and Mitra method introduced earlier), and a third where only the special uniques are considered. Informally, a statistical unique can occur by either chance (random unique) or it can occur because of a rare combination of traits (special unique). Special uniques are deemed more risky than random uniques; Elliot et al. (2002) define an algorithm for scoring statistical uniques according to how special they are.

To identify statistical uniques and special uniques, we used the Special Uniques Detection Algorithm (SUDA) software (Elliot et al. 2002). For each statistical unique, SUDA generated a score using the Data Intrusion Simulation (DIS) method, which estimates the intruder confidence in a match leading to correct inference (see Elliot et al. 2002, for more details). We used the records in the top decile of scores generated by SUDA so as to examine the most risky of records (the special uniques). For the LCF there were 1867 statistically unique records and 251 special unique records, while for the BSA there were 2120 statistically unique records and 235 special unique records.

4 Results and Discussion

The CAP scores for the LCF dataset are shown in Table 1[6]. It shows that for the synthetic datasets all CAP scores are smaller than the CAP score for the

[6] The different imputation levels (m) are nested, rather than independent synthetic datasets.

Table 1. Mean CAP scores for the original and synthetic LCF datasets for two methods handling non-matches, two synthesis methods, three different intruder scenarios; full set, statistical uniques, and special uniques and ten levels of multiple imputation.

	Non-matches as zero			Non-matches as undefined		
Scenario	Full set	Statistical uniques	Special uniques	Full set	Statistical uniques	Special uniques
Original	0.750	1	1	0.750	1	1
Baseline	0.266	0.255	0.226	0.266	0.255	0.226
CART						
$m = 1$	0.334	0.180	0.074	0.498	0.548	0.549
2	0.393	0.273	0.110	0.503	0.554	0.530
3	0.416	0.324	0.154	0.501	0.549	0.568
4	0.435	0.361	0.162	0.505	0.545	0.535
5	0.443	0.380	0.176	0.502	0.537	0.525
6	0.448	0.388	0.192	0.501	0.532	0.525
7	0.453	0.397	0.212	0.500	0.524	0.522
8	0.459	0.411	0.218	0.502	0.529	0.507
9	0.463	0.421	0.242	0.502	0.529	0.523
10	0.465	0.427	0.242	0.502	0.528	0.519
Parametric						
$m = 1$	0.296	0.138	0.0418	0.459	0.433	0.525
2	0.346	0.208	0.0531	0.460	0.433	0.430
3	0.364	0.251	0.0774	0.452	0.435	0.485
4	0.378	0.277	0.0817	0.450	0.434	0.437
5	0.388	0.295	0.0920	0.449	0.431	0.436
6	0.393	0.304	0.101	0.447	0.426	0.408
7	0.397	0.315	0.0987	0.445	0.427	0.393
8	0.403	0.324	0.108	0.447	0.428	0.394
9	0.406	0.328	0.115	0.446	0.424	0.384
10	0.420	0.355	0.142	0.455	0.437	0.410

original data. When the non-matches are coded as undefined all datasets have a CAP that is larger than the baseline CAP, there is no substantial effect of the number of imputations. The differences between the CAP scores for the three scenarios (full set, statistical uniques, and special uniques) are inconsistent but not large.

On the other-hand, when the non-matches are coded as zero, the full set, statistical uniques, and special uniques have different CAP sizes, with the CAP size becoming smaller as the records become riskier. Additionally, as m increases the CAP score increases. We found for the full set that the CAP scores were larger than the baseline CAP score. However, for the special uniques, the synthetic CAP score was similar to or smaller than the baseline CAP scores, but as m increases the synthetic CAP score becomes larger than the baseline CAP score.

This relationship between the synthetic CAP scores and the baseline CAP score is different than the findings of Elliot (2014), who found that the synthetic and baseline had similar scores. This difference most likely stems from two factors: (1) Elliot was using a smaller key size. Appendix A shows that when smaller keys are observed the difference between the synthetic CAP score, the original CAP score, and baseline CAP score is less dramatic. (2) Elliot was only using single imputation. As seen in Table 1 as m increases so does the CAP score, hence a dataset that is only $m = 1$, would be less disclosive and therefore closer to the baseline CAP score.

Additionally, when the non-matches are coded as zero, Table 1 shows that the statistical unique CAP scores tend to be smaller than the full set CAP scores, and the special unique scores are smaller than the statistical unique scores. This trend is not so when the non-matches are undefined. The statistical uniques and special uniques are actually a bit larger than the full set. This indicates that while riskier records, as designated by the statistical uniques and special uniques, are less likely to have a match, if they do have a match it is just as likely to be correct as any other match.

Table 1 also shows that for the synthetic CAP scores the parametrically generated synthetic dataset had smaller CAP scores than the CART generated synthetic dataset. This shows that - in this experiment at least - the parametrically generated synthetic data has less risk than the CART generated synthetic data.

4.1 CAP Scores Regressed on Number of Imputations

To explore the relationship between the CAP scores and the number of imputations (m), we put the CAP scores into a simple linear regression analysis where y is the CAP score and m is a continuous variable, shown in Table 2. Table 2 shows that when the non-matches are coded as zero (Models 1–3) the number of imputations (m) has a significant and positive effect. When the non-matches are coded as undefined, the regression models show a different relationship. Table 2 shows that when the non-matches are coded as undefined, for the full set (Model 4), the relationship between m and the CAP score is not significant for CART, but has a significant, if small, negative coefficient for the parametric synthetic data. For model 5 when the data is parametric the relationship is not significant, but has a small negative coefficient for the CART data. For the special uniques (Model 6) there is not significant relationship between m and the CAP score. The m coefficient for Models 4, 5 and 6, is, even when significant, considerably smaller than the m coefficient for Models 1, 2, or 3.

When the non-matches are coded as CAP=0, m has a significant positive effect on the CAP score, when non-matches are coded as undefined for the CAP score there is essentially no effect. The increase when non-matches are zero, is an artefact of the lower number of non-matches. With more synthetic samples, non-matches are less likely to occur and this increases the CAP score. However, when the non-matches are undefined there is no reason for the CAP score to

Table 2. Simple linear regression of CAP score on the number of imputations - LCF

	Term	Non-matches as zero			Non-matches as undefined		
		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
		Full set	Statistical uniques	Special uniques	Full set	Statistical uniques	Special uniques
CART	Intercept	0.367***	0.221***	0.0809***	0.501***	0.548***	0.552***
	m	0.0117***	0.0234***	0.0177***	5.64e-05	−0.00283*	−0.00393
Parametric	Intercept	0.367***	0.169***	0.0431***	0.458***	0.438***	0.489***
	m	0.0117***	0.0186***	0.00838***	−0.00141*	−0.00165	−0.0114

*p < 0.05 **p < 0.01 ***p < 0.001

change in either direction, hence m has either a very small coefficient and mostly non-significant coefficient.

4.2 Comparing the LCF CAP Scores to the BSA CAP Scores

The results for the BSA dataset mostly confirm the results from the exploratory regression for the LCF. (The average CAP scores for the BSA can be found in Appendix B). Table 3 is an exploratory regression analysis of the BSA CAP scores. Like the LCF, when the non-matches are coded as zero (Models 1, 2, and 3) m has a significant positive relationship to the CAP score, showing that as m increases the likelihood of a match increases. However, for the CAP scores when the non-matches are excluded the BSA is different than the LCF. For the full set (Model 4) for the parametric synthetic data m has a significant negative coefficient. While for Model 5 the statistical uniques have significant m coefficients Model 4 probably has a significant coefficient since the BSA has a larger proportion of statistical uniques than the LCF (see Sect. 3.3) and therefore Model 4 will look more similar to Model 5. Additionally, Appendix B shows that like the LCF, for the BSA the CART CAP scores are larger than the parametric CAP scores.

Table 3. Simple linear regression of CAP score on the number of imputations - BSA

	Term	Non-matches as zero			Non-matches as undefined		
		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
		Full set	Statistical uniques	Special uniques	Full set	Statistical Uniques	Special uniques
CART	Intercept	0.143***	0.114***	0.0109	0.302**	0.114***	0.337***
	m	0.0137***	0.0157***	0.0201***	0.000517	0.0157***	0.00908
Parametric	Intercept	0.0615***	0.0420***	0.00335	0.177***	0.169***	0.220**
	m	0.00641***	0.00677***	0.00445***	−0.00226**	−0.00219*	0.00539

*p < 0.05 **p < 0.01 ***p < 0.001

4.3 Is Synthetic Data Differentially Confidential?

Here we introduce the notion of *differential confidentiality* as an alternative way to using the CAP score to the DCAP method described in Sect. 2.1. We determine that a dataset is differentially confidential in respect of a given target and key if on average there is no difference in the CAP score for a record whether the record is in the original dataset or not.

To demonstrate this concept for synthetic data we partitioned the data into two equal sized datasets and synthesised one (A) but not the other(B). We then calculated the synthetic CAP score for both A and B. Table 4 shows that synthetic data is not inherently differentially confidential, since if a record is included in the synthesis model it has a higher average CAP score than were it not included in the synthesis model. It is noteworthy that the differences between the two sets are larger for the CART synthetic data than it is for the parametric synthetic data. Also of note, for the t-tests [7], while all t-statistics for the CART synthetic data were significant at the $p < 0.001$ level, for the parametric synthetic data, the t-statistics were significant at the $p < 0.05$ level if they were significant at all, which some were not. Furthermore, as m increases the CART synthetic data becomes less differentially confidential, while that is not the case for the parametrically generated synthetic data.

The reader may have noted a superficial similarity of this concept with *differential privacy*. However, this is a post hoc test where as the former is a mechanism for achieving a standard[8]. It is possible that a dataset could be differentially private but not differentially confidential and vice versa.

4.4 Summary

With respect to DCAP the greater number of imputations, the more likely a match is to occur. However the likelihood of said match leading to a correct inference is not affected by the number of imputations, this picture can only fully be seen by looking at both the CAP scores when non-matches are coded as zero and undefined. In all instances the synthetic CAP score is lower than that of the original, showing that the synthetic data does decrease the risk of disclosure. While the disclosure risk is less than that of the original, it does not satisfy differential confidentiality.

That being said, synthetic data does particularly decrease the disclosure risk of special uniques, which are the most risky records in any microdata. The special uniques had CAP scores at the same level or lower than the baseline CAP, showing that the risk to special uniques from synthetic data is less than releasing a univariate distribution. In all scenarios, the CART synthetic data had a higher CAP score than the parametric synthetic data. This all suggests that parametrically synthetic data has less disclosure risk than CART synthetic data.

[7] We used Welch's T-Test DF = 5,131.

[8] See for example Abowd and Vilhuber, 2008; Charest 2010 for uses of differential privacy in the synthesizing mechanism.

Table 4. Results of the differential confidentiality test for the LCF synthetic dataset-with non-matches as zero.

	Included in synthesis	Not included in synthesis	Difference	T-test
CART				
$m = 1$	0.295	0.207	0.0875	8.0987***
2	0.370	0.252	0.1181	10.835***
3	0.404	0.269	0.134	12.514***
4	0.426	0.282	0.144	13.553***
5	0.442	0.294	0.148	14.13***
6	0.455	0.300	0.155	14.85***
7	0.462	0.303	0.160	15.503***
8	0.473	0.312	0.161	15.691***
9	0.476	0.315	0.162	15.841***
10	0.481	0.319	0.161	15.871***
Parametric				
$m = 1$	0.253	0.231	0.0215	2.0214*
2	0.318	0.289	0.0295	2.6972**
3	0.341	0.319	0.0216	1.9972*
4	0.361	0.339	0.0213	1.9825*
5	0.372	0.347	0.0248	2.3498*
6	0.376	0.350	0.0258	2.4819*
7	0.384	0.361	0.0234	2.271*
8	0.386	0.368	0.0186	1.8221
9	0.391	0.375	0.0156	1.5404
10	0.397	0.378	0.0185	1.8377

5 Conclusion

In this paper we have developed the methods introduced by Elliot (2014) for measuring attribute disclosure risk in synthetic data. The CAP score appears to be a simple but robust measure of attribute disclosure risk and the two approaches to using that measure - DCAP and differential confidentiality - seem to provide some traction on the difficult problem of measuring attribute disclosure. Indeed, we note that although these methods have been developed with a view to measuring the disclosure risk for synthetic data, they could be used on any structured datasets. Indeed, they might be useful for the calculation of the relative residual risk of different disclosure control and privacy protection methods. In future work we hope to develop such a general comparative methodology.

A An exploration into the CAP scores when smaller sized keys are used or the LCF

Key 6: GOR, Output area classifier, tenure, dwelling type, internet in hh, household size
Key 5: GOR, Output area classifier, tenure, dwelling type, internet in hh
Key 4: GOR, Output area classifier, tenure, dwelling type
Key 3: GOR, Output area classifier, tenure (Table 5).

Table 5. Mean CAP scores for the original and synthetic LCF datasets for for two methods of handling non-matches, two synthesis methods, three different key sizes, three different intruder scenarios;and ten levels of multiple imputation.

Scenario	Full set			Statistical uniques			Special uniques		
File	Key 5	Key 4	Key 3	Key 5	Key 4	Key 3	Key 5	Key 4	Key 3
Original	0.610	0.560	0.466	1	1	1	1	1	1
Baseline	0.266	0.266	0.266	0.244	0.248	0.239	0.242	0.246	0.239
Non-matches as zero									
CART									
$m = 1$	0.378	0.381	0.395	0.200	0.196	0.234	0.109	0.159	0.234
2	0.411	0.401	0.402	0.289	0.275	0.292	0.147	0.191	0.292
3	0.422	0.411	0.404	0.327	0.324	0.338	0.169	0.233	0.338
4	0.428	0.414	0.406	0.350	0.333	0.343	0.186	0.229	0.343
5	0.431	0.416	0.406	0.367	0.352	0.345	0.191	0.239	0.345
6	0.433	0.418	0.406	0.373	0.363	0.351	0.202	0.239	0.351
7	0.434	0.419	0.406	0.375	0.367	0.355	0.202	0.243	0.355
8	0.437	0.421	0.407	0.381	0.371	0.351	0.216	0.265	0.351
9	0.437	0.421	0.407	0.380	0.373	0.352	0.240	0.279	0.352
10	0.439	0.422	0.408	0.385	0.376	0.358	0.249	0.289	0.358
Parametric									
$m = 1$	0.360	0.362	0.388	0.153	0.152	0.197	0.0727	0.0685	0.197
2	0.387	0.382	0.396	0.217	0.203	0.231	0.0720	0.0929	0.231
3	0.395	0.383	0.393	0.242	0.221	0.282	0.0988	0.147	0.282
4	0.400	0.388	0.395	0.263	0.238	0.288	0.0874	0.137	0.288
5	0.404	0.390	0.396	0.279	0.255	0.297	0.0908	0.151	0.297
6	0.405	0.391	0.395	0.290	0.267	0.307	0.124	0.162	0.307
7	0.406	0.392	0.395	0.297	0.279	0.302	0.146	0.193	0.302
8	0.406	0.392	0.395	0.296	0.274	0.289	0.166	0.206	0.289
9	0.408	0.392	0.394	0.309	0.278	0.289	0.188	0.234	0.289
10	0.415	0.398	0.396	0.328	0.306	0.308	0.214	0.278	0.308
Non-matches as undefined									
CART									
$m = 1$	0.444	0.421	0.404	0.466	0.441	0.432	0.632	0.702	0.432
2	0.448	0.422	0.405	0.478	0.437	0.358	0.523	0.518	0.358
3	0.449	0.425	0.406	0.471	0.44	0.376	0.476	0.502	0.376
4	0.450	0.427	0.406	0.469	0.441	0.365	0.445	0.458	0.365
5	0.449	0.427	0.407	0.472	0.445	0.364	0.428	0.455	0.364
6	0.449	0.427	0.407	0.467	0.443	0.370	0.436	0.436	0.370
7	0.449	0.427	0.407	0.460	0.440	0.371	0.419	0.425	0.371
8	0.450	0.428	0.407	0.457	0.431	0.360	0.417	0.427	0.360
9	0.449	0.427	0.407	0.449	0.425	0.358	0.433	0.427	0.358
10	0.449	0.428	0.408	0.448	0.425	0.361	0.441	0.434	0.361

(*continued*)

Table 5. (*continued*)

Scenario	Full set			Statistical uniques			Special uniques		
File	Key 5	Key 4	Key 3	Key 5	Key 4	Key 3	Key 5	Key 4	Key 3
Original	0.610	0.560	0.466	1	1	1	1	1	1
Baseline	0.266	0.266	0.266	0.244	0.248	0.239	0.242	0.246	0.239
Parametric									
$m = 1$	0.428	0.405	0.397	0.373	0.324	0.318	0.471	0.338	0.318
2	0.429	0.406	0.400	0.373	0.320	0.287	0.377	0.325	0.287
3	0.425	0.400	0.395	0.372	0.309	0.320	0.402	0.398	0.320
4	0.424	0.401	0.396	0.374	0.312	0.314	0.332	0.319	0.314
5	0.425	0.400	0.396	0.376	0.315	0.310	0.322	0.334	0.310
6	0.423	0.399	0.396	0.376	0.318	0.318	0.367	0.325	0.318
7	0.422	0.399	0.395	0.375	0.323	0.310	0.382	0.330	0.310
8	0.421	0.398	0.395	0.365	0.312	0.293	0.381	0.327	0.293
9	0.421	0.398	0.395	0.370	0.314	0.294	0.414	0.352	0.294
10	0.425	0.402	0.396	0.382	0.333	0.308	0.429	0.390	0.308

B The average CAP scores for the BSA

(See Table 6).

Table 6. Mean CAP scores for the original and synthetic BSA datasets for two methods handling non-matches, two synthesis methods, three different intruder scenarios; full set, statistical uniques, and special uniques and ten levels of multiple imputation.

Scenario	Non-matches as zero			Non-matches as undefined		
	Full set	Statistical uniques	Special uniques	Full set	Statistical uniques	Special uniques
Original	0.876	1	1	0.876	1	1
Baseline	0.0853	0.0851	0.0869	0.0853	0.0851	0.0869
CART						
$m = 1$	0.115	0.0871	0.0127	0.291	0.304	0.273
2	0.173	0.146	0.0503	0.311	0.329	0.370
3	0.196	0.172	0.0631	0.302	0.313	0.309
4	0.218	0.197	0.0971	0.306	0.316	0.362
5	0.230	0.211	0.130	0.307	0.316	0.424
6	0.238	0.221	0.149	0.304	0.312	0.417
7	0.248	0.233	0.160	0.308	0.316	0.409
8	0.251	0.238	0.173	0.306	0.314	0.419
9	0.257	0.246	0.187	0.307	0.315	0.414
10	0.259	0.248	0.190	0.305	0.312	0.407
Parametric						
$m = 1$	0.0474	0.0309	0.00426	0.168	0.154	0.200
2	0.0716	0.0517	0.0128	0.168	0.159	0.273
3	0.0896	0.0701	0.0213	0.176	0.171	0.263
4	0.0998	0.0812	0.0241	0.173	0.170	0.258
5	0.101	0.0829	0.0241	0.161	0.156	0.227
6	0.106	0.0875	0.0281	0.161	0.153	0.220
7	0.111	0.0943	0.0338	0.163	0.156	0.240
8	0.113	0.0959	0.0380	0.159	0.151	0.255
9	0.114	0.0974	0.0416	0.155	0.148	0.271
10	0.115	0.100	0.0506	0.153	0.147	0.297

References

Caiola, G., Reiter, J.: Random forests for generating partially synthetic, categorical data. Trans. Data Priv. **3**, 27–42 (2010)

Charest, A.: How can we analyze differentially-private synthetic datasets? J. Priv. Confid. **2**(2), 21–33 (2010)

Chen, Y., Elliot, M., Sakshaug, J.: A genetic algorithm approach to synthetic data production. In: PrAISe 2016 Proceedings of the 1st International Workshop on AI for Privacy and Security, Hague, Netherlands. ACM (2016)

Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: Principles of Database Systems, pp. 202–210 (2003)

Drechsler, J.: Using support vector machines for generating synthetic datasets. In: Domingo-Ferrer, J., Magkos, E. (eds.) PSD 2010. LNCS, vol. 6344, pp. 148–161. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15838-4_14

Drechsler, J., Bender, S., Rässler, S.: Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel. Trans. Data Priv. **1**, 105–130 (2008)

Drechsler, J., Reiter, J.: An empirical evaluation of easily implemented, non-parametric methods for generating synthetic data. Comput. Stat. Data Anal. **55**, 3232–3243 (2011)

Elliot, M.: Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team. CMIST (2014).http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/. Accessed 17 Mar 2017

Elliot, M., Mackey, E., O'Hara, K., Tudor, C.: The Anonymisation Decision-Making Framework, 1st edn. UKAN, Manchester (2016)

Elliot, M., Mackey, E., O'Shea, S., Tudor, C., Spicer, K.: End user licence to open government data? A simulated penetration attack on two social survey datasets. J. Off. Stat. **32**(2), 329–348 (2016)

Elliot, M., Manning, A., Ford, R.: A computational algorithm for handling the special uniques problem. Int. J. Uncerta. Fuzziness Knowl.-Based Syst. **10**(5), 493–509 (2002)

Fienberg, S., Makov, U.: Confidentiality, uniqueness, and disclosure limitation for categorical data. J. Off. Stat. **14**(4), 385–397 (1998)

Kim, J., Winkler, W.: Masking microdata files. In: Proceedings of the Survey Research Methods Section, pp. 114–119. American Statistical Association (1995)

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

Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: L-diversity: privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data **1**(1), 1–52 (2007)

NatCen Social Research: British Social Attitudes Survey, 2014, [data collection], UK Data Service, 2nd edn (2016). Accessed 30 Apr 2018. SN: 7809. https://doi.org/10.5255/UKDA-SN-7809-2

Nowok, B., Raab, G., Dibben, C.: synthpop: Bespoke creation of synthetic data in R. J. Stat. Softw. **74**(11), 1–26 (2016)

Office for National Statistics: Department for Environment, Food and Rural Affairs, Living Costs and Food Survey, 2014, [data collection], UK Data Service, 2nd edn (2016). Accessed 08 Mar 2018. SN: 7992. https://doi.org/10.5255/UKDA-SN-7992-3

Raab, G., Nowok, B., Dibben, C.: Practical data synthesis for large samples. J. Priv. Confid. **7**(3), 67–97 (2016)

Reiter, J.: Using CART to generate partially synthetic, public use microdata. J. Off. Stat. **21**, 441–462 (2005)

Reiter, J., Mitra, R.: Estimating risks of identification disclosure in partially synthetic data. J. Priv. Confid. **1**(1), 99–110 (2009)

Reiter, J., Wang, Q., Zhang, B.: Bayesian estimation of disclosure risks for multiply imputed, synthetic data. J. Priv. Confid. **6**(1), 17–33 (2014)

Rubin, D.B.: Statistical disclosure limitation. J. Off. Stat. **9**(2), 461–468 (1993)

Skinner, C., Elliot, M.: A measure of disclosure risk for microdata. J. R. Stat. Soc. Ser. B **64**(4), 855–867 (2002)

Smith, D., Elliot, M.: A measure of disclosure risk for tables of counts. Trans. Data Priv. **1**(1), 34–52 (2008)

Winkler, W.: Re-identification methods for evaluating the confidentiality of analytically valid microdata. Research Report Series, 9 (2005)

Yancey, W., Winkler, W., Creecy, R.: Disclosure risk assessment in perturbative microdata protection. Research Report Series, 1 (2002)