Practical privacy metrics for synthetic data

Gillian M Raab, Beata Nowok & Chris Dibben

December 15, 2024

Abstract

This paper explains how synthpop package [1] [2] has been extended to include functions to calculate measures of identity and attribute disclosure risk for synthetic data that measure risks for the records used to create the synthetic data. The basic function `disclosure` calculates identity disclosure for a set of quasi-identifiers (keys) and attribute disclosure for one variable specified as a target from the same set of keys. The second function `multi.disclosure` is a wrapper for the first and presents summary results for a set of targets. This short paper explains the measures of disclosure risk and documents how they are calculated. We recommend two measures: *RepU* (replicated uniques) for identity disclosure and *DiSCO* (Disclosive in Synthetic Correct Original) for attribute disclosure. Both are expressed a % of the original records and each can be compared to similar measures calculated from the original data. Experience with using the functions on real data found that some apparent disclosures could be identified as coming from relationships in the data that would be expected to be known to anyone familiar with its features. We flag cases when this seems to have occurred and provide means of excluding them.

1 Introduction

In his recent review of synthetic data (SD) methodology Reiter [17] comments:

> "While there is need to examine disclosure risks in synthetic data, there is no standard for doing so, especially in fully synthetic data. Instead, disclosure checks tend to be ad hoc"

[1] see `www.synthpop.org.uk`

[2] The new version 1.8-1 is currently available on Github at `https://github.com/gillian-raab/synthpop`

This is in contrast to the variety of measures of utility available for SD; see [16] for a list of these. While utility measures must be chosen that are relevant to the intended use of the SD, disclosure measures must focus on the possible harm to the privacy of an individual or other unit whose data contributed to the creation of the SD. The evaluation of the disclosure risk from SD must relate to the context of its release (see [6] for a discussion of this). We cannot expect a fixed rule, for example that a criterion for release requires a value of some disclosure measure beyond a threshold. Instead, we expect those releasing data to use the disclosure measures to evaluate potential harm to the data subjects, or to the data custodians, from information in the released data. We hope that these new functions will allow the disclosure risk of SD sets to be explored and, where necessary, reduced.

We expect that disclosure risks from SD to be lower than those from the original data. But in some cases, e.g. a large data set with only a few categorical variables, the disclosure risk from the original may already be low. To evaluate the protection from disclosure afforded by a synthesis method, the risks for SD must be compared with equivalent risks for the GT. We consider two types of disclosure risk:

- *identity disclosure*: This refers to the ability to identify individuals in the data from a set of known characteristics that we will refer to as keys. Identity disclosure may be less relevant for completely synthesized[3] data because there is no one-to-one correspondence between records in the original and SD. But it may still be of interest since it is an important factor for attribute disclosure.

- *attribute disclosure*:This refers to the ability to find out from the keys something, not previously known, for an attribute associated with a record in the original data.

The measures we describe here are appropriate for fully SD where all items in all records are replaced by synthetic values. Different measures have been developed for partially SD [3, 18]. Note that our disclosure methods treat numeric variables, by default, as if they were categories unless `ngroups_keys` and/or `ngroups_target(s)` are set. In our first example the `income` variable has been grouped into 20 categories, but the other numerical variable (depress) with only 21 categories has been left as it is. The number of groups formed may differ from the parameter setting e.g. if

[3]Complete or full synthesis is when all values of all variables are replaced by synthetic values. This is in contrast to incomplete or partial synthesis where only some variables are replaced.

there are fewer than `numgroups` distinct values.[4].

The disclosure risk posed by SD can be reduced by using techniques from statistical disclosure control (**sdc**), such as aggregation of categories, smoothing of numeric values or removal of replicated uniques. These methods can be used to reduce disclosure risk by modifying the original data before synthesis, or the SD before its is released. Some such methods are already available in the **synthpop** package. These include categorising, top/bottom coding and smoothing for continuous variables, and the merging of small categories for factors. The removal of replicated uniques is another option available in `synthpop`.

There have been many recent proposals for making synthetic data sets comply with Differential Privacy (DP) [4]. DP is a very strong privacy guarantee that protects against an intruder with arbitrary external information about the subjects in the data, except for the one whose privacy is being protected. This is an unrealistic assumption and DP SD has been shown to have unacceptably low utility in many cases [1, 20]. We will not discuss these methods here, but note that we could use the metrics proposed here to evaluate disclosure risks for DP SD.

2 A simple example

Here we illustrate the basic use of `multi.disclosure`. If the parameter `targets` is not specified, all the variables in the SD that are not part of keys are used as targets. The identity disclosure measures are `UiO` for original and `repU` for synthetic, and for attribute disclosure `Dorig` for original and `DiSCO` for synthetic. These and other measures will be explained in Sections 3.2 and 3.3.

First, a subset of 9 variables are selected from the `SD2011` data (a survey on quality of life in Poland) that is available as part of the `synthpop` package. A single synthetic data set is created by the default method in `synthpop`: `cart` for each conditional distribution. Note that the variable `income` has values -8 that indicate not applicable, and the synthesis allows for this. The synthetic data object `s1`[5] has a component `syn` that is a single synthetic data set. The disclosure functions can also be used with synthetic data created by other methods either as single synthetic data sets or lists of repeated syntheses from the same original. Here we select 4 keys that represent items that might be known for members of this sample, or of the Polish population in 2011. By default,the 5 other variables become the targets: `depress income ls marital` (marital status),

[4]The code uses grouping options from functions in the `classInt` package.

[5]an object of class `synds`

workab (intention to work abroad). The second target, income, is grouped into 20 categories, plus the -8 category. We first use multi.disclosure to create a multi.disclosure object for these keys, and we print out its identity component.

```
R> library("synthpop")
R> ods <- SD2011[, c("sex", "age", "region","placesize","depress",
+    "income","ls","marital" , "workab")]
R> s1 <- syn(ods, seed = 8564, print.flag = FALSE, cont.na = list(income = -8))
R> t1 <- multi.disclosure(s1, ods, print.flag = FALSE, plot = FALSE,
+  keys = c("sex", "age", "region", "placesize"),   ngroups_targets = c(0,20,0,0,0))
R> print(t1,to.print = "ident")

Disclosure risk for 5000 records in the original data

Identity disclosure measures
from keys: sex age region placesize
For original  ( UiO )  48.38 %
For synthetic ( repU ) 14.86 %.
```

The measure UiO (Unique in Original) shows that 48% of the original records would have unique combinations of these 4 keys. The term, "singling out" is used in data protection regulation for this type of attribute disclosure[6]. For the synthetic data $RepU$ tells us that almost 15% of the original records would be unique in the original and also unique in the synthetic data. For attribute disclosure we can examine the results as either a table or a plot. Here we see in Figure 1 the plot that would have been generated if plot had been set to TRUE in the code above, the attribute disclosure results would have been plotted as shown in Figure 1. Details of how identity and disclosure measures are calculated can be found in Section 3. The results in the Figure 1 are in descending order by the attribute disclosure risk in the SD.

The measure $Dorig$ tells us that these 4 keys would identify a unique value of each of these targets for over 50% of the original records. We get lower values for $DiSCO$, the proportion of the original records that are disclosive in the original and also in the synthetic data with a correct attribution to the target. The first two variables shown in Figure 1 have additional labels that flag possible contributions to disclosure from knowledge of 1-way or 2-way relationships in the original data. This is detailed in Section 4.

Previous work on attribute disclosure [5, 22] has used the Correct Attribution Probability ($DCAP$) as a disclosure measure for synthetic data. This is calculated as the percentage of records

[6]See for example its use in the UK Information Commissioner's guidance on anonymisation here https://ico.org.uk/media/about-the-ico/documents/4018606/chapter-2-anonymisation-draft.pdf

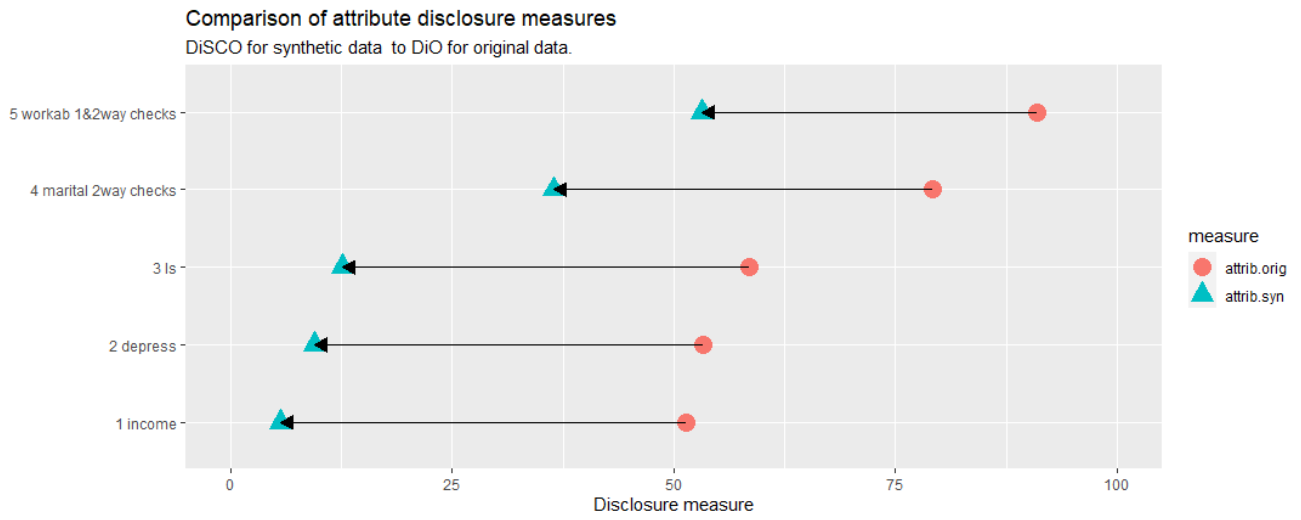

Figure 1: Plot from `multi.disclosure` with `plot = TRUE`, the default value.

correctly predicted from the synthetic data by a set of keys. However it has been suggested [2], that it may be as much a measure of utility as disclosure risk and modifications have been suggested [10, 21] one of which ($TCAP$) is close to our proposed measure, $DiSCO$. For completeness the measures $DCAP$ and $TCAP$ are calculated by `disclosure`; see Appendix 2.

3 Scenario and definitions

3.1 Setting the scene

These disclosure measures are intended to assess what a person who only has access to the synthetic data can infer about known individuals who are present in the original data. We will use the term "intruder" for such a person, though no malicious intent is implied. The intruder is assumed to have information for one or more individuals about the value of certain key variables that are present in the same format in the original and the synthetic data. They attempt first to see if the individual is present, and then to determine the value of other items in the data file that we refer to as targets. We are assuming a worst-case scenario where the intruder believes they are querying the original data.[7] Disclosure measures from the synthetic data are each compared to similar measures for someone with access to the original data. Here we will introduce the measures by an example. Formal definitions with notation and formulae are in Appendix 1. The first step

[7] This may not be too unrealistic if the data are made available inadvertently, or if the intruder thinks that efforts to label the synthetic data as e.g. "Fake Data" are thought just to be a cover up. It may also be a reqasonable measure to use without this scenario, since it compares the disclosiveness of the SD to that of the GT.

in evaluating disclosure risk, as described here, is to identify a set of keys that might be expected to be known to an intruder. These keys are then combined to form a quasi-identifier that we designate as q. For example, if we have hospital records we might define age,sex date and hospital as keys and this would give a q with levels such as "`78 | M | 1/1/2024 | WG`" for a 78 year old man admitted to hospital WG on 1/1/2024.

3.2 Identity disclosure measures

The concept of k-anonymity is central to identity disclosure for microdata. First proposed in 1998 [19] it is discussed fully in [6]. A table is k-anonymous if a set of keys identifies at most k individuals. Thus 2-anonymous data will never identify just one individual. Based on this idea, the percentage of records for which the keys identify just one individual give identity disclosure measures. Tables of q values are produced from the synthetic and the original data. UiO and UiS are the percentages of records with keys where the table count is 1. An intruder checking out a record for their known set of keys will look for it in the synthetic data. Some records will not be in the SD and $UiOiS$ (Unique in Original in Synthetic) gives the % that would be found. These records are then checked for uniqueness in the synthetic data giving, *repU* is the percentage of unique original records that are also unique in the SD.

The percentage *repU* has been used as a disclosure measure to evaluate SD by [9] and by [15] [8]. Replicated uniques are used in `synthpop` as part of the statistical disclosure control function, `sdc`, that includes the option of reducing disclosure risk by removing them from the SD. Nowok et al. [13] have evaluated this and give an example where this process has very little effect on utility. The function `replicated.uniques`[9] also calculates *repU* using a different method from the one described here.

One of the outputs of the functions `disclosure` and `multi.disclosure` is `ident`, a table of identity disclosure measures as illustrated by this example, using the keys from the example in Section 2 but now calculated for a synthetic object with 5 data sets, using the one target `depress`. We first create `t5` an object of class `disclosure` and then print out the identity and attribute disclosure measures for each synthetic data set.

The table for identity disclosure has UiO and *repU* as its first and last column. The 2nd and

[8] Jackson et al. in [9] argue that the denominator for *repU* should be N_s rather than N_d. This is inappropriate because our scenario is to consider the risk to the original data.

[9] For example by `replicated.uniques (s2, ods, keys = c("sex","region","age","placesize")`

3rd columns are UiS calculated from the synthetic data in the same manner as UiO from the original and $UiOiS$ the % of UiO with q that are in the SD but not necessarily unique. Each are steps towards calculating $repU$.

```
R> s5 <- syn(ods, seed = 8564, m = 5, print.flag = FALSE)
R> t5 <- disclosure( s5, ods, keys = c("sex", "age", "region",
+    "placesize"), target = "depress", print.flag = FALSE)
R> print(t5, to.print = c("ident"))

Disclosure measures from synthesis for 5000 records in original data.

Identity disclosure measures for 5 synthetic data set(s) from keys:
 sex age region placesize

    UiO   UiS UiOiS  repU
1 48.38 37.34 22.68 14.86
2 48.38 35.44 22.24 13.96
3 48.38 35.18 21.98 13.62
4 48.38 34.90 22.08 13.78
5 48.38 36.14 22.00 14.62
```

3.3 Attribute disclosure measures

We approach disclosure risk from the point of view of an intruder with access to the SD and to certain attributes (quasi-identifiers) known for one or more individuals in the original data. The quasi-identifiers for the known record(s) can be combined to create a composite variable q that can be created for all records in the original and SD. Modelling what an intruder might do we calculate the following measures, each of which is a proportion of the original records:

- Look up q for each original record in the SD. The proportion found becomes iS (in Synthetic).
- Check if all records with the same q have the same level of the target t. The proportion passing this further test becomes DiS (Disclosive in Synthetic).
- Then check if these apparent disclosures correspond to the value of t in the original data. The proportion of original records for which this is true becomes $DiSCO$ Disclosive in Synthetic Correct Original.

Note that records contributing to $DiSCO$ may not be disclosive in the original data, as this information would not be available to the intruder. A further measure $DiSDiO$ (Disclosive in

Synthetic Disclosive in Original) restricts the score to those disclosive in the original. These measures are defined formally in Appendix 1.

```
R> print(t5, to.print = c("attrib"))

Disclosure measures from synthesis for 5000 records in original data.

Attribute disclosure measures for depress from keys: sex age region placesize
  Dorig  Dsyn    iS   DiS DiSCO DiSDiO max_denom mean_denom
1  53.3 46.26 64.90 34.18  9.54   6.14         3       1.16
2  53.3 44.80 64.00 32.50 10.26   6.78         4       1.19
3  53.3 44.60 64.02 32.14  9.10   5.92         4       1.19
4  53.3 45.80 63.88 33.38  9.20   5.52         4       1.21
5  53.3 44.52 63.44 31.46  9.34   5.80         4       1.23
```

Here we print the table of attribute disclosure measures for the example in the previous section. As we move from iS to DiS and to $DiSCO$ we can see how the %disclosive is affected by different conditions. Here lack of q levels in the SD retains just 64% of records. The requirement for a record to be disclosive in the SD reduces this to 34% and again to around 9% for those with a correct attribution. This reduces it to around 6% by restricting to records disclosive in the original.

The $Dorig$ and $DiSCO$ measures are not restricted to disclosures that are identified from unique records for q in either the original or the SD. The number of records contributing to each disclosive qt cell in the synthetic table is the denominator that applies to that record. The columns `max denom` and `mean denom` refer to the denominators for the records disclosive in the SD that contribute to $DiSCO$. We can see from the mean that here the majority of disclosive records had unique key combinations in the SD, and the maxima was 3 for the first synthesis and 4 for the others. Large denominators can be an indication that some of the disclosure may be coming from strong relationships between variables in the data that might even be expected a-priori. This aspect is discussed further in Section 4. The disclosure measures can be restricted to those with small denominators by using the parameter `exclude_over_denom_lim` to TRUE. The example in Section 2 is here run restricted to denominators of 1.

```
R> multi.disclosure(s1, ods, print.flag = FALSE, plot = FALSE,
+    keys = c("sex", "age", "region", "placesize"),
+    denom_lim =1, exclude_ov_denom_lim = TRUE)

Disclosure risk for 5000 records in the original data

Identity disclosure measures
from keys: sex age region placesize
```

```
For original  ( UiO )  50.26 %
For synthetic ( repU ) 16.36 %.

Table of attribute disclosure measures for sex age region placesize
Original measure is  Dorig and synthetic measure is DiSCO
Variables Ordered by synthetic disclosure measure

                     attrib.orig attrib.syn                    check1 Npairs
1 income                   48.38       3.36                                0
2 depress                  48.38       7.02                                0
3 ls                       48.38       9.84                                0
4 marital                  48.38      15.40                                0
5 workab 1way checks       48.38      20.48 Check  workab  level  NO      0
                     check2
1 income
2 depress
3 ls
4 marital
5 workab 1way checks
```

The *DiSCO* values have decreased, as expected, but *UiO* and *repU* have increased and also that *Dorig* is now the same for all targets and equal to the *UiO* value before the denominator exclusion. This makes sense because removing large denominators from *UiO* and *repU* increases the number of uniques. Also, with denominators of 1, all unique values of q are disclosive in the original data.

Note that decreases in *DiSCO* are more pronounced for `workab` and `marital`, the two targets that were flagged to check 1-way or 2-way relationships. Their original *DiSCO* values were 37% for `marital` and 53% for `workab`. This and other methods of excluding certain apparent disclosures are discussed in the next section.

4 Identifying disclosure from 1-way and 2-way relationships

As mentioned in the Introduction, what we can learn about disclosiveness of attributes can depend on our prior knowledge of the data set or the population from which it is drawn. It would not be practical to specify our prior probability for every possible combination of keys. However, checking two aspects of disclosure results can help us to check when we might have predicted the correct attribution with high probability without knowledge of all the keys. The first is a check for a target where a high proportion of records have one level of the target. The second is when

there is a strong relationship between a target and one of the keys, so that one tq pair accounts for many of the disclosive records. These two aspects are flagged by the values of `check_1way` and `check_2way` that are returned as part of objects of Returning to the example in Section 2, we now use the function `disclosure` to get details for the two targets that were flagged as requiring checking; see Figure 1.

```
R> d1_workab <- disclosure(s1, ods, print.flag = FALSE, target = "workab",
+     keys = c("sex", "age", "region", "placesize"),plot = FALSE)
R> print(d1_workab, to.print = c("check_1way"))

Disclosure measures from synthesis for 5000 records in original data.

Details of target level contributing disproportionately to disclosure
   Level  All PctLevelAll totalDisclosive nLevelDis PctLevelDis
NO    NO 5000       88.64            2605      2482       95.28

Details of target-key pairs contributing disproportionately to disclosure of workab
26 pairs need checks

Please examine component $check_2way of the disclosure object and look at original data.
Consider excluding these key-target pairs with some the following parameters to disclosure
exclude_ov_denom_lim = TRUE or defining key-target combinations from exclude.targetlevs,
exclude.keys and exclude.keylevs
```

We can see that it was the category "NO" of `workab` that contributed most to the disclosure risk; most survey respondents (89%) had never worked abroad. This level of the target accounted for 95% of apparent disclosures for `workab`. To predict this level for all of a group with the same q could hardly be considered disclosive. This would be true if the intruder had access to the marginal distribution of `workab`, and even if they did not, some knowledge of the respondents' background might suffice. The tq pairs identified for `workab` both included the "NO" level of the target.

```
R> d1_marital <- disclosure(s1, ods, print.flag = FALSE, target = "marital",
+     keys = c("sex", "age", "region", "placesize"),plot = FALSE)
R> print(d1_marital, to.print = c("check_2way"))

Disclosure measures from synthesis for 5000 records in original data.

Details of target-key combinations contributing disproportionately to disclosure
This is a list with one for each of 4 6 syntheses
  target_key_levs npairs key key_target_total key_total PctTargetKeyLevel
7       SINGLE|19     10 age              91        92             98.91
4     MARRIED|41      5 age              59        73             80.82
6       SINGLE|18      5 age              91        92             98.91
8       SINGLE|22      5 age              85        91             93.41
```

For the target `marital` two tq pairs are identified as contributing large denominators. Those aged 19 are almost all SINGLE, and those aged 40 are mainly MARRIED. Again these would not be considered disclosive for an intruder with some knowledge of the respondents.

The thresholds for identifying one-way and two-way relationships can be modified. In each case there are two criteria that need to be satisfied, one number and one %. For one-way disclosure the parameter `thresh_1way` has the default value of $c(50, 90)$, meaning that there must be at least 50 disclosive records for one level of a target and that q values including this target must account for over 90% of all disclosive records. For two way relationships the `thresh_2way` has default value $c(5, 80)$ and the algorithm first identifies all disclosive tq combinations with denominators over 4. It then identifies the level of the key in q with that best predicts this level of t and checks if over 80% of the disclosive records with this key would have the correct prediction.

5 Excluding records

Having identified records where the apparent disclosure is something that would generally be known, one option is to exclude these key-target combinations explicitly from the measures. The following parameters for `multi.disclosure` and `disclosure` can be used for this.

- `not.target`: All records with levels of the target given by the parameter `not.target` are excluded from all disclosure measures.
- `usekeysNA`:This is set to TRUE by default so that missing values are included in all tables. It can be set to FALSE for some or all keys to exclude NAs.
- `usetargetsNA)(multi.disclosure) or usetargetNA (disclosure) )`: Similar to the above for target(s) - note can be a vector for multi.disclosure.
- `exclude.keys, exclude.keylevs and exclude.targetlev`: Three vectors of length the number of key-target pairs to be excluded from all tables.[10]

To illustrate exclusions we will use the Adult data set from the UCI machine learning repository [12] with almost 50 thousand records from the US Census income study, available as part of the `arules` package for **R** [8]. This is one of the data sets used in [7] to evaluate privacy risks for SD.

[10]For `multi.disclosure` these parameters are supplied as lists with elements for each target in `targets`

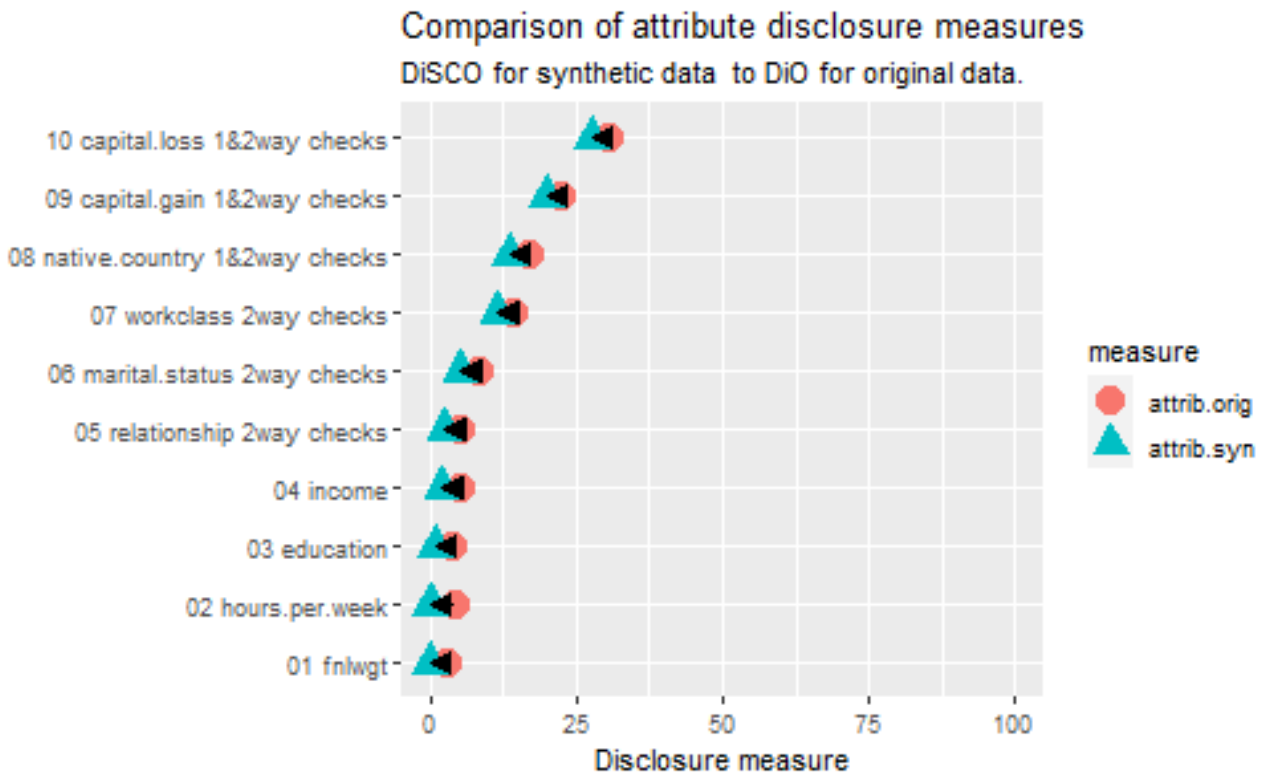

Figure 2: Plot from `multi.disclosure` for variables flagged by checks.

Note that the two variables, `capital gain` and texttcapital gain, originally numeric data, have been truncated so that they have only 123 and 99 distinct values repsectively. They have not been grouped further in the results presented here. Figure 2 is the output of `multi.disclosure` from the keys `age, occupation, race and sex` for the other 10 variables that are not in the keys. The disclosiveness of the original data is relatievly low, compared to the previous example, because of the large sample size and the absence of any geographic identifiers. Three of the 10 variables `capital.gain, capital.loss and native.country` are flagged to check one-way relationships. The check1 column of the attribute table tells us that the levels contributing to disclosure are zeros for the first two and `United-States` for `native.country`. These levels make up 95%, 92% and 89% of all records.

Three other variables are flagged as having disclosive two_way relationships `workclass, marital, relationship` with totals of 7,7 and 2 tq combinations respectively. The summary function `multi.disclosure` does not allow target and key specific pairs to be excluded. To investigate this it is necessary to examine the output of `disclosure`[11]. This showed that the largest

[11] see code in Appendix 3 for how to do this.

contribution to `check_2way` for `workclass` was due to this being missing when occupation was missing, although other relationships between these two variables also contributed.

	No excl		not.tlev		NAs out not.tlev		denom_lim 1 NAs out not.tlev		denom_lim 1	
target	orig	syn	orig	syn	orig	syn	orig	syn	orig	syn
capital.gain	22.55	19.87	0.21	0.00	0.21	0.00	0.19	0.00	2.68	1.17
capital.loss	30.61	27.71	0.08	0.00	0.08	0.00	0.08	0.00	2.68	1.26
education	3.71	0.93	3.71	0.93	3.71	0.93	2.45	0.46	2.68	0.50
fnlwgt	2.70	0.00	2.70	0.00	2.70	0.00	2.45	0.00	2.68	0.00
hours.per.week	4.36	0.19	4.36	0.19	4.36	0.19	2.45	0.12	2.68	0.13
income	4.97	2.10	4.97	2.10	4.97	2.10	1.58	0.50	2.68	0.82
marital.status	8.23	5.27	8.23	5.27	8.23	5.27	2.45	0.68	2.68	0.74
native.country	17.09	13.55	0.94	0.08	0.94	0.08	0.67	0.04	2.68	0.87
relationship	5.17	2.64	5.17	2.64	5.17	2.64	2.45	0.69	2.68	0.73
workclass	14.27	11.64	14.27	11.64	14.27	11.64	2.45	0.85	2.68	0.96

Table 1: Disclosure results from Adult data with different exclusions.

Table 1 gives the results of excluding different entries in the tables of q and t from the attribute disclosure measures. Excluding the levels of the target flagged by `check_1way` reduces the $DiSCO$ to almost zero for these 3 variables. Adding exclusion of missing values reduced the disclosure for `workclass` and for some other variables a little. Adding the restriction to denominators of 1, reduced the disclosure for variables identified by `check_2way` to low levels. In the final columns we can see that restricting to denominators of 1, by itself, gives low levels of disclosure for all variables, with the exception of those flagged by `check_1way`. This approach was one that was trialed in an earlier version of these functions but discarded for giving misleading results for some data. It does not count some attribute disclosures that might well be found by an intruder. For example, if a small number of records with the same keys all had the same level of the target that corresponded to the level in the original data. We feel that a better approach is to exclude specific target/key combinations.

6 Conclusions

The privacy metrics we propose here are in some sense the opposite of differential privacy (DP). DP claims to protect data from an intruder with arbitrary knowledge of the data, except of the one record that has the greatest influence on the likelihood of the results. In contrast our metrics

require the user to specify keys that identify variables in the data that would expect to be known about individuals, as well as to specify the details of what they would expect an intruder to know about the data. Also, the routines flag cases where part of the disclosure measures come from one- or two-way relationships so that disclosure would be expected even in the absence of data.

We have evaluated the routines on a few data sets and set levels of the thresholds for this at what seem to be reasonable levels, but more experience with other data sets would be valuable. We hope that these tools will be helpful to data holders who need to make decisions about the risks of releasing SD to the public, or to a restricted audience. They should also enable the disclosure risks of different synthesis methods to be evaluated.

7 Acknowledgement

Research Data Scotland (`https://www.researchdata.scot/` funded Gillian Raab's time to carry out the research reported here and to expand the capabilities of the synthpop package[12] to include measures of disclosure control. We also thank the Scottish Centre for Administrative Data Research for continue to support the development of the synthpop package since its creation was supported by the ESRC funded SYLLS poject in 2012-14.

References

[1] Bowen, C. M., and Liu, F. Comparative study of differentially private data synthesis methods. *Statistical Science 35* (2020), 280—-307.

[2] Chen, Y., Taub, J., and Elliot, M. Trade-off between information utility and disclosure risk in ga synthetic data generator. UNECE Work Session on Statistical Data Confidentiality, Skpoje, North Macedonia. Available from `https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Chen_Taub_Elliot_AD.pdf`, 2019. Accessed: 2024-01-06.

[3] Drechsler, J., and Reiter, J. P. Disclosure risk and data utility for partially synthetic data: An empirical study using the german iab establishment survey. *Journal of official statistics 25*, 4 (2009), 589–603.

[12] from version 1.8-0 available on CRAN at `https://CRAN.R-project.org/package=synthpop`

[4] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 24th Annual International Conference on The Theory and Applications of Cryptographic Techniques* (2006), EUROCRYPT'06, pp. 486–503.

[5] Elliot, M. Final report on the disclosure risk associated with the synthetic data, produced by the sylls team. Available from `https://tinyurl.com/syllsDR`, 2014. Accessed: 2022-02-23.

[6] Elliot, M., Mackey, E., and O'Hara, K. The anonymisation decision-making framework: European practitioners. Available from `https://ukanon.net/framework/`, 2020. Accessed: 2022-02-23.

[7] Giomi, M., Boenisch, F., Wehmeyer, C., and Tasnádi, B. A unified framework for quantifying privacy risk in synthetic data, 2022.

[8] Hahsler, M., Gruen, B., and Hornik, K. arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software 14*, 15 (October 2005), 1–25.

[9] Jackson, J., Mitra, R., Francis, B., and Dove, I. Using saturated count models for user-friendly synthesis of large confidential administrative databases. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 185* (2022), 1613—-1643.

[10] Little, C., Elliot, M., and Allmendinger, R. Comparing the utility and disclosure risk of synthetic data with samples of microdata. In *Privacy in Statistical Databases* (Cham, 2022), J. Domingo-Ferrer and M. Laurent, Eds., Springer International Publishing, pp. 234–249.

[11] Machanavajjhala, A., Gehrke, J., Kifer, K., and Venkitasubramaniam, M. l-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)* (2006), IEEE.

[12] Newman, C. B. D., and Merz, C. UCI repository of machine learning databases, 1998.

[13] Nowok, B., Raab, G., and Dibben, C. Recognising real people in synthetic microdata: risk mitigation and impact on utility. UNECE Work Session on Statistical Data Confidentiality, Skpoje, North Macedonia. Available from https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/3_risk_mitigation.pdf, 2017. Accessed: 2022-02-23.

[14] Pater, L., and Smid, S. Making attribute information of synthetic data interpretable with the aggregation equivalence level. UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS Expert Meeting on Statistical Data Confidentiality 26-28 September 2023, Wiesbaden. Available from https://unece.org/sites/default/files/2023-08/SDC2023_S4_3_Netherlands_Pater_D.pdf, 2024. Accessed: 2022-02-23.

[15] Raab, G. Utility and disclosure risk for differentially private synthetic categorical data. In *Privacy in Statistical Databases, 2022* (2022), K. Muralidhar and J. Domingo-Ferrer, Eds., Springer, Berlin.

[16] Raab, G. M., Nowok, B., and Dibben, C. Assessing, visualizing and improving the utility of synthetic data, 2021.

[17] Reiter, J. Synthetic data: A look back and a look forward. *Transactions in Data Privacy 16* (2023), 15—-24.

[18] Reiter, J. P., and Mitra, R. Estimating risks of identification disclosure in partially synthetic data. *The journal of privacy and confidentiality 1*, 1 (2009).

[19] Samarati, P., and Sweeney., L. Generalizing data to provide anonymity when disclosing information. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* (1998), ACM Press.

[20] Stadler, T., Oprisanu, B., and Troncoso, C. Synthetic data – anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)* (Boston, MA, 2022), pp. 1451–1468.

[21] Taub, J., and Elliot, M. The synthetic data challenge. UNECE Work Session on Statistical Data Confidentiality,29-31 October 2019, the Hague, the Netherlands. Available from https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/

SDC2019_S3_UK_Synthethic_Data_Challenge_Elliot_AD.pdf, 2019. Accessed: 2024-06-12.

[22] Taub, J., Elliot, M., Pampaka, M., and Smith, D. Differential correct attribution probability for synthetic data: An exploration. In *Privacy in Statistical Databases - UNESCO Chair in Data Privacy, International Conference, PSD 2018, Proceedings* (2018), F. Montes and J. Domingo-Ferrer, Eds., Springer-Verlag Italia, pp. 122–137.

Appendix 1: Notation and formal definitions

Before defining the measures of identity and disclosure risk we need to introduce the notation that will be used to calculate them. The first step is to create the quasi-identifiers from the keys for the original and SD. For the keys used in the example given in Section 2 the quasi-identifier that we will designate as q for the first record in the original data is:

`"FEMALE | 57 | Lubuskie | URBAN 100,000-200,000"`

and that for the first record in the SD:

`"FEMALE | 39 | Zachodnio-pomorskie | URBAN 100,000-200,000"`.

In order to calculate identity disclosure measures, we need to compare the tables of q from the original and SD. For attribute disclosure measures we need to cross-tabulate q with each target variable t and compare findings from the SD with what would have been found from the original data. In general, the levels of q and sometimes t in the original and SD will not be the same. Before creating any tables, we need to define sets of q and t values that give the union of both sets of levels and align the tables so that their indices correspond.

For the original data $d_{.q}$ is the count of records with the keys corresponding to the levels of q and q_{tq} is the count of records with this q and level $t = 1, ...T$ of the target. The equivalent counts from the synthesised data are designated by $s_{.q}$ and s_{tq}. When a member of q is in the original data but not in the synthetic, $s_{.q}$ and s_{tq} are all zero. Similarly when a member of q is in the SD but not in the original, $d_{.q}$ and d_{tq} are all zero. The two tables can be written as shown in Table 1, where the total records in the original data is N_d, made up of $N_{d\ only}$ and $N_{d\ both}$. The equivalent totals for the SD are N_s, $N_{s\ only}$ and $N_{s\ both}$.

	only in original	in both	only in synthetic	Total
1	... d_{1q} ...	... d_{1q} ...	... 0 ...	$d_{1.}$
...	... ...	... ...	... ...	...
t	... d_{tq} ...	... d_{tq} ...	... 0 ...	$d_{t.}$
...	... ...	... ...	... ,,, ...	...
T	... d_{Tq} ...	... d_{Tq} ...	... 0 ...	$d_{T.}$
Column sums	$d_{.q}$	... $d_{.q}$ ...	... 0 ...	N_d
Totals	$N_{d\ only}$	$N_{d\ both}$	0	N_d

	only in original	in both	only in synthetic	Total
1	... 0 ...	... s_1q ...	... s_1q ...	$s_{1.}$
...	... ...	... ...	... ...	...
t	... 0 ...	... s_tq ...	... s_tq ...	$s_{t.}$
...	... ...	... ...	... ...	...
T	... 0 ...	... s_Tq ...	... s_Tq ...	$s_{T.}$
Column sums	... 0 ...	... $s_{.q}$ ...	... $s_{.q}$ ...	N_s
Totals	0	$N_{s\ both}$	$N_{s\ only}$	N_s

Table 2: Notation for tables from quasi-identifier (q) and target (t) from original (upper table) and SD (lower table).

To calculate the % of records in the original and SD we need:

$$\%\ Unique\ in\ Original = UiO = 100 \sum (d_{.q} | d_{.q} = 1)/N_d. \tag{1}$$

$$\%\ Unique\ in\ Synthetic = UiS = 100 \sum (s_{.q} | d_{.q} = 1)/N_d. \tag{2}$$

The intruder has information about the keys for an individual in the real data that they attempt to identify in the SD. They first attempt to find them in the SD, and the % found is:

$$\%\ Unique\ in\ Original\ in\ Synthetic = UiOiS = 100 \sum (d_{.q} = 1 | s_{.q} = 1 \wedge d_{.q} > 0)/N_d. \tag{3}$$

Some of these records would not be unique in the SD, restricting to such records gives:

$$\%\ replicated\ Uniques = repU = 100 \sum (s_{.q} | d_{.q} = 1 \wedge s_{.q} = 1)/N_d. \tag{4}$$

To find an attribute from a set of keys, it is necessary to examine the distribution of s_{tq} for groups defined by q. We define column proportions for the original and SD as $pd_{tq} = d_{tq}/d_{.q}$ and for the synthetic as $ps_{tq} = s_{tq}/s_{.q}$.

Returning to the scenario described in Section 3.1, we must first define a measure of attribute

disclosure for the original data. This is based on the concept of *l-diversity* [11] that requires that each set of records defined by q has at least $l(\geq 2$ distinct values of the target. A data set is *l2-diverse* for q and t if all records for every q have the same level of t[13] An attribute disclosure measure for the original data can be defined as *% Disclosive in Original* :

$$Dorig = 100 \sum^{q} \sum^{t} (d_{tq}|pd_{tq} = 1)/N_d. \tag{5}$$

The equivalent measure for the SD, taken as if it were real, becomes :

$$Dsyn = 100 \sum^{q} \sum^{t} (s_{tq}|ps_{tq} = 1)/N_s. \tag{6}$$

An intruder with access only to the SD, but with knowledge of q from one or more individuals in the original, would look them up in the SD. Some of their q levels be key combinations that do not appear in the SD leaving the proportion that do appear as iSO (in in Synthetic Original)

$$iSO = 100 \sum^{q} \sum^{t} (d_{tq}|s_{tq} > 0)/N_d. \tag{7}$$

A level of q from an original record may identify more than one target in the SD, or identify the wrong target. To exclude these we require the records to be Disclosive in the SD and Correct when checked with the original giving:

$$DiSCO = 100 \sum^{q} \sum^{t} (d_{tq}|ps_{tq} = 1)/N_d. \tag{8}$$

Note that $DiSCO$ can include records that are not disclosive in the original data giving a further measure Disclosive in Synthetic and Disclosive in the Original :

$$DiSDiO = 100 \sum^{q} \sum^{t} (d_{tq}|ps_{tq} = 1 \wedge pd_{tq} = 1)/N_d. \tag{9}$$

As we comment above the intruder would not be able to tell if records were identified as $DiSDiO$ rather than $DiSCO$, so we prefer the latter measure. However, the intruder can identify when the apparently disclosive record is not unique in the SD. This restriction can be imposed by

[13]This could be generalised to $l > 2$, but in practice the levels of targets are not generally exchangeable and a more practical approach would be to aggregate levels for certain targets.

requiring that the denominator for disclosive records in the SD does not exceed a 1, as described and discussed in Section 5.

Appendix 2: CAP measures

The measures *baseCAPd*, *DCAP*, *TCAP* are calculated by the function `disclosure` and are stored in the component `allCAPs` of the output object of class `disclosure`. This is printed when the parameter `to.print` includes `allCAPs`. The first measure is known as the baseline CAP and refers to an average of the predictions that would be made by someone who only has access to the marginal distribution of the target. The intruder then guesses the CAP for each level of the target according to the relative frequencies $pd_{t.}$. Averaging this over all observations gives

$$baseCAPd = \sum (pd_{t.})^2 / N_d.$$

The % of t that would be correctly predicted from q for someone with access to the original data is $CAPd$:

$$CAPd = \sum_{tq} (pd_{tq} d_{tq}) / N_d,$$

and the equivalent measure for the SD treated as if it were the original is

$$CAPs = \sum_{tq} (ps_{tq} s_{tq}) / N_s.$$

The measure DCAP is the percentage of t correctly predicted in the SD q. This gives

$$DCAP = \sum_{tq} (ps_{tq} d_{tq}) / N_d$$

The $DCAP$ score can be expressed by scaling it from $baseCAPD$ to 1 (see [10, 14]), although this can result in some negative values. The $TCAP$ measure as defined by [10] uses the same count of disclosive records as $DISCO$, but uses as denominator the number of records that have keys represented in both the synthetic and the original data. It can also be scaled from $baseCAPD$ to 1, and as [10] demonstrate this measure can also take negative values.

```
R> print(t5, to.print = "allCAPs")
```

```
Disclosure measures from synthesis for 5000 records in original data.

CAP measures for the target depress with keys
sex age region placesize

  baseCAPd  CAPd  CAPs  DCAP  TCAP
1     9.81 74.15 69.78 16.39 14.70
2     9.81 74.15 69.36 17.45 16.03
3     9.81 74.15 69.24 16.20 14.21
4     9.81 74.15 69.87 15.92 14.40
5     9.81 74.15 69.00 16.17 14.72
```

As expected the lower denominator for $TCAP$ gives values higher than the equivalent $DiSCO$ values for the same example given in Section 3.3.