On the privacy guarantees of synthetic data: a reassessment from the maximum-knowledge attacker perspective

Nicolas Ruiz[1], Krishnamurty Muralidhar[2], Josep Domingo-Ferrer[3]

[1,3]Universitat Rovira i Virgili
Department of Computer Science and Mathematics
CYBERCAT-Center for Cybersecurity Research of Catalonia
UNESCO Chair in Data Privacy
Av. Països Catalans 26, 43007 Tarragona, Catalonia
nicolas.ruiz@oecd.org;josep.domingo@urv.cat

[2] University of Oklahoma
Department of Marketing & Supply Chain Management
Price College of Business
308 Brooks Street, Norman OK 73019, USA
krishm@ou.edu

Abstract. Generating synthetic data for the dissemination of individual information in a privacy-preserving way is an approach that is often presented as superior to other statistical disclosure control techniques. The reason for such claim is straightforward at first glance: since all records disseminated are synthetic and not actual observed values, no individual can reasonably claim to face a privacy threat. Thus, and if the synthesizer used is good enough, synthetic data will potentially always offer a high level of information with low disclosure risk attached. Building on recent advances in the literature regarding the conceptualization of an intruder, this paper aims at challenging this claim by reassessing the privacy guarantees of synthetic data. Using the concept of a maximum-knowledge intruder, we demonstrate that synthetic data can in fact be always expressed as a re-arrangement of the original data and that, as a result, they may lead to configurations where disclosure risk may be higher than for non-synthetic disclosure control approaches. We illustrate the application of these results by an empirical example.

Keywords. Statistical disclosure control, Synthetic data, Maximum-knowledge attacker

1 Introduction

Data on individual subjects are increasingly collected and exchanged. By their nature, they provide a rich amount of information that can inform statistical and policy

analysis in a meaningful way. However, due to the legal obligations surrounding these data, this wealth of information is often not fully exploited in order to protect the confidentiality of respondents. In fact, such requirements shape the dissemination policy of microdata at national and international levels. The issue is how to ensure a sufficient level of data protection to meet releasers' concerns in terms of legal and ethical requirements, while offering users a reasonable richness of information. Moreover, over the last decade the role of microdata has changed from being the preserve of National Statistical Offices and government departments to being a vital tool for a wide range of analysts trying to understand both social and economic phenomena. As a result, more parties, often very heterogeneous in their privacy and information requirements, are now involved in microdata transactions. This has opened a new range of questions and pressing needs about the privacy/information trade-off and the quest for best practices that can be both useful to users but also respectful of respondents' privacy.

Statistical disclosure control (SDC) research has a rich history in addressing those issues, by providing the analytical apparatus through which the privacy/information trade-off can be assessed and implemented. SDC consists in the set of tools that can enhance the level of confidentiality of any data while preserving to a lesser or greater extent their level of information (see [7] for an authoritative survey). Over the years, it has burgeoned in many directions. In particular, techniques applicable to microdata, which are the focus of this paper, offer a wide variety of tools to protect the confidentiality of respondents while maximizing the information content of the data released, for the benefits of society at large.

While generally considered as part of the SDC literature, the publication of synthetic data is an appealing alternative to, but also a significant departure from, pure SDC methods. The idea is simple: instead of disseminating an anonymized version of a dataset, i.e. the original data altered by the application of an SDC method, some data are instead created by drawing from a model fitted to the original data (hereafter called a synthesizer). At first glance it is clear that, since all values are synthetic and none of the individuals in the original data are included, disclosure risk must be practically non-existent [14]. The original data are used to build the synthesizer, and thus the contribution of an individual to a data set is not pointless but is in fact used only as an informational basis. As a result, synthetic data seem to offer a clear and almost definitive advantage compared to other SDC methods: it would seem that synthetic data can be made as close as possible to the original data without any strong concern for privacy, while for non-synthetic SDC methods similarity to original data must be traded off against disclosure risk in a more stringent way (and hence utility is necessarily limited).

However, further scrutiny appears to weaken the advantage offered by synthetic data. For the sake of illustration, assume a dystopian society in possession of a perfect synthesizer, i.e. one that is able to perfectly replicate the statistical information observed over its population. In this case, an intruder using the synthetic data to conduct his attack may be able to re-identify some individuals or learn some sensitive information about them. From the point of view of the individuals, the fact that the information gained by the intruder is synthetic does not change much the situation: the

right to privacy has been violated. While from a legal perspective this situation may not be unlawful [19], from an ethical perspective this can be clearly qualified as a negative outcome. Of course, in real life the perfect synthesizer does not exist. But the better the job done by the data releaser to create the synthetic data, the closer can be an attacker to gaining valuable information about some respondents in the original data. Thus, it can be reasonably argued that, ultimately, synthetic data are somehow subject to the same kind of risk/information trade-off faced by non-synthetic SDC methods.

It is based on these considerations that the privacy guarantees of synthetic data need to be explicitly considered. In [18], a privacy model to produce synthetic data with *ex ante* privacy guarantees was proposed. Here, we take an *ex post* approach, as previously performed in e.g. [6,12], but based on a new, encompassing definition of an attacker, in order to reassess the privacy guarantees of synthetic data, regardless of how they have been obtained. The definition of an attacker on individual data has always been a thorny issue in the literature, not least because one must postulate how much background knowledge the attacker has. As a result, a variety of scenarios can be constructed, all based on *ad hoc* assumptions that may not comply with the views and the constraints faced by the data releasers, and that will also remain very context-specific. A recent proposal in the SDC literature has tried to circumvent these difficulties by proposing the concept of a maximum-knowledge attacker [1]. This attacker is based on a rather radical setting because he is entitled with the knowledge of *both* the original and the anonymized data set. While this may appear as unrealistic at first glance, such a scenario is conceptually powerful: any anonymized data set judged as sufficiently safe in term of disclosure risk under this scenario will in fact be safe under any kind of other possible scenario. As a result, this concept is a way to unify the comparison of the various performances of SDC techniques by using a common benchmark. It is also a way to ease the dissemination of individual data in the sense that, if a data releaser agrees with the level of disclosure risk contained in his anonymized data set under the maximum-knowledge attacker configuration, then he can be reassured that the release will be safe whatever the malicious attempts that could take place on his data.

Now, if the notion of a maximum-knowledge attacker seems valuable to gauge non-synthetic SDC techniques, it seems fair to submit synthetic data to the same kind of test. This is the purpose of this paper, structured as follows. Section 2 gives some background concepts on synthetic data and the maximum-knowledge attacker model needed later on. Section 3 characterizes the consequences of having some synthetic data submitted to a maximum-knowledge attack, and subsequently derives some new tools to assess their privacy guarantees. Section 4 presents some empirical results based on these tools. Conclusions and future research directions are gathered in Section 5.

2 Background concepts

2.1 Synthetic data

Synthetic data rely on a principle that is by nature similar to the imputation of missing values in a data set. The idea is to fit a model, called a synthesizer, to the original data; then values are drawn from the synthesizer to replace original data rather than merely imputing missing data. Three types of synthetic data can be distinguished [7]:

- *Fully synthetic data*: no original data are released and the values of all attributes across all records are synthetic.
- *Partially synthetic data*: across some if not all records, only sensitive attributes are synthesized while for example quasi-identifiers are original values.
- *Hybrid data*: original and fully synthetic data are combined, and the resulting data can be more or less similar to the original or fully synthetic data.

The above distinction will not have any consequences in what follows in this paper, so we will use the term synthetic data indistinctively to point to any of the three types. However, what is common to them is obviously the pivotal role of the synthesizer. Generating synthetic data worth disseminating is work-intensive, not least because creating a synthesizer that can replicate the intricate features of a micro data set necessitates some time and an involved level of expertise. It is beyond the scope of this paper to discuss the relative merits of the several approaches available to create a synthesizer, as well as the criteria that can be used to gauge it (see [4] for an extensive discussion), but a general principle is that the level of information offered by a synthetic data set can be only as good as the quality of the underlying synthesizer used to generate it. In this paper, we will simply assume that the data releaser did a good enough job so that the resulting synthetic data are worth disseminating and being analyzed by the users.

Regarding the practical characteristics of synthetic data, let us emphasize that they do not always come under the same format than the original data. First of all, they do not have to be of the same size, although having the same number of synthetic records than the number of original records seems a natural choice. To the best of the authors' knowledge, no firm guideline exists in the literature on this criterion (see however [13] for an empirical discussion). Depending on the context, an argument can be made for releasing synthetic data smaller than, same size as, or larger than the original data. Given this, we will assume that the number of synthetic records is the same as the original data. However, we will not restrict to the case of equal number of synthetic and original records, as one of the appeals of synthetic data is that they can come under any size. Specifically, we will outline below a pre-sampling procedure that can be applied before undertaking the evaluation of the privacy guarantees of synthetic data; this will in fact allow gauging synthetic data sets of any size.

A second difference with non-synthetic SDC methods is that synthetic data generally lead to the dissemination of several data sets, while for the former methods only one set is released. This practice is motivated by the goal of capturing the different

designs of the original data [15]. Clearly, such a feature can quickly become cumbersome for the users (as well as for the releasers who need to generate the sets under various design configurations) and thus has to balance cost and accuracy [13]. Moreover, in the case where the original data are numerical and approximately multivariate normal, the sufficiency-based perturbation approach will perform at least as well as synthetic data for the preservation of information, while at the same time necessitating the release of only a single data set, which eases the tasks of the users [10].

Here again, no firm guideline exists on the right number of data sets to be released. The original proposal of releasing multiple data sets postulates as a rule-of-thumb a typical number between 3 and 10 [15], but later contributions outlined that this number is in fact context-dependent and may vary according to the analytical needs of the user and the properties of the employed synthesizer [13]. In this paper, we will assume that an arbitrary number M of synthetic data sets is released. As we will demonstrate, this number will turn out to be critical for the privacy guarantees of synthetic data.

Finally, in the introduction of this paper we briefly touched upon the fact that disclosure risk in fully synthetic data must always be by nature almost non-existent. Such claim has been made at various occasions in the literature, e.g. [4,5,13,14], albeit it must be mentioned that: i) this conclusion is less clear-cut for partially synthetic or hybrid data [5,7] (which by construction will contain some of the original data), ii) as far as pure synthetic data are concerned, some attempts to evaluate disclosure risk have also been previously proposed [6,12]. In these last two cases however, it is again generally assumed that the risk is very low. The recent advances in the SDC literature on the notion of intruder cast a new light on this crucial feature of synthetic data.

2.2 The maximum-knowledge attacker model

The issue of an attacker's background knowledge has been recently pushed further in the literature through the proposal of the maximum-knowledge attacker model [1,9]. This model defines an attacker who knows both the original data set and its entire corresponding anonymized version. This is a rather extreme configuration, unlikely to be mirrored by concrete situations, but it remains however conceptually very insightful, as an anonymized data set that can pass the test of such a situation will in fact be able to pass any test. It also has the advantage of completely resolving the issue of which background knowledge is to be assumed for operationalizing an attack on individual data [9]. Moreover, this model has as a consequence that, while it legitimates an exclusive focus on re-identification disclosure, it can be easily adapted to attribute disclosure assessment by excluding from the maximum-knowledge background a specific attribute [2]. In that case, the attacker's objective is to learn about the specific attribute's values as precisely as possible. This possibility, of particular interest for synthetic data, will be exploited in the empirical section of this paper.

The concept of a maximum-knowledge attacker is rooted in the known-plaintext attack defined in cryptology. While other types of attack can be conceived, they carry less meaning in the context of individual data [1]. A ciphertext-only attack, where only the anonymized data set is available, is less realistic than a known-plaintext at-

tack: the attacker is likely to know at least a few original records or attributes, as part of his background knowledge. Regarding chosen-plaintext or chosen ciphertext-attacks, they are relevant only in the case in which the attacker can interact with the anonymization procedure; this may occur when protecting the answers to interactive queries to an on-line database, but it does not happen when releasing an anonymized data set. Note that a maximum-knowledge attacker, observing both the original data set and its anonymized version, has nothing to gain in terms of information. One can view his attempt as being purely slanderous, trying to discredit the data releaser by revealing his anonymization procedure.

Given the assumption that such a powerful person might exist, this leads to one question: what is exactly the perspective of that intruder? In fact, the reply relies on the record tracking numbers. Generally, and after having applied a non-synthetic SDC technique, data releasers can track which anonymized record derives from which original record through a number that does not carry any information of any sort and is unaffected by the anonymization procedure. Moreover, when the data are released, all numbers can be modified or deleted. But these numbers, known for practical purposes by the data releaser but not by the maximum-knowledge attacker, act in fact as a mask. Contrary to the statement made in [16], record tracking numbers in fact set the limit of the maximum-knowledge attacker. To make this clear, Table 1 illustrates the attacker's perspective on a toy example. In this example, the intruder has to retrieve the mapping between records in *X* and records in *Y*. His task is equivalent to retrieving some permutation structures. Note that permutation is in fact the overarching principle governing non-synthetic SDC methods [1,11,16,17]. In what follows, we will also demonstrate that actually such principle applies to synthetic methods too.

Table 1. Point of view of a maximum-knowledge attacker

Original dataset X					Anonymized dataset Y			
ID	X_1	X_2	X_3		ID	Y_1	Y_2	Y_3
1	13	135	3707			8	160	3248
2	20	52	826			20	57	822
3	2	123	-1317			-1	122	248
4	15	165	2419			18	135	597
5	29	160	-1008			29	164	-1927

Using synthetic data does have some implications for the maximum-knowledge attacker model. For non-synthetic SDC methods, the releaser has the advantage over the maximum-knowledge attacker of knowing the mapping between the tracking numbers in *X* and *Y*. The releaser can use this knowledge for example to assess how an individual has been protected; even the individual herself can verify her protection, if she can identify her own record in the non-synthetic data set. But *for synthetic methods the mapping between original and synthetic records does not make much sense*: a synthetic record does not derive from any specific single original record. Thus, *the advantage of the releaser over the maximum-knowledge attacker vanishes: both are at the same level of knowledge*. The privacy risk in synthetic data is not tied to a mapping: it is rather connected with knowing that synthetic records exist that are very close to some original records. In fact, real and synthetic individuals are linked by

information. This can be assessed by a multivariate version of a rank-based record linkage procedure that will be developed below.

3 Synthetic data from the maximum-knowledge attacker perspective

3.1 Multiple reverse mapping of synthetic data

We first start by observing that a synthetic data releaser can always transform the data such that each attribute in each synthetic data set can be expressed as a permutation of the original data. This procedure, called reverse mapping, has been recently proposed in the literature for non-synthetic SDC methods [1,11]. To the best of our knowledge, this is the first time that it is developed for synthetic data.

Assume that a releaser generates $m = 1, \dots, M$ synthetic data sets $Y^m = (Y_1^m, \dots, Y_p^m)$ based on an original data set $X = (X_1, \dots, X_p)$; denote by $X_j = (x_{1,j}, \dots, x_{n,j})$ and $Y_j^m = (y_{1,j}^m, \dots, y_{n_m,j}^m)$ the values of attribute $j=1,\dots,p$ over n records in the original data and n_m records in the m^{th} synthetic data set, respectively. No further assumptions are made, except that the values of an attribute can always be ranked, which is obvious in the case of numerical or categorical attributes, but also feasible in the case of nominal ones [3].

In particular, the synthetic data sets need not be of the same size as the original data set. However, in order to perform reverse mapping, we need to compare sets of the same size. This issue can be fixed as follows: when the synthetic data have more (resp. less) records than the original data, synthetic data can be randomly sub-sampled (resp. super-sampled):

- When $n_m > n$, a subset Q^m of size n is randomly selected;
- When $n_m < n$, a superset Q^m of size n is created by randomly generating n-n' additional records from the original n' ones;
- When $n_m = n$, the synthetic data are not modified and $Q^m = Y^m$.

Such a preliminary sampling procedure is viable provided that the original data set is large enough for it to be analytically interesting and representative. In the remainder of this paper, we will assume that $n_m = n, \forall m = 1, \dots, M$, keeping in mind that the pre-sampling procedure can be eventually used to align the sizes of every synthetic data sets with the size of the original data. The multiple reverse mapping of synthetic data is then performed as follows:

***Algorithm**: multiple reverse mapping of synthetic data*
***Require**: original data set X, with attributes $X_j = (x_{1,j}, \dots, x_{n,j})$, for j=1,...,p*
***Require**: synthetic data sets Y^m, for m=1,..., M, where Y^m has attributes $Y_j^m = (y_{1,j}^m, \dots, y_{n,j}^m)$, for j=1,...,p*
For** m=1, ..., M **do
    For** j=1,...,p **do

For *i=1,…,n* ***do***
Compute k=Rank($y_{i,j}^m$)
Set $z_{i,j}^m = x_{(k,j)}$ (where $x_{(k,j)}$ is the value of X_j of rank k)
Next *i*
Let $Z_j^m = (z_{1,j}^m, \dots, z_{n,j}^m)$
Next *j*
Let *data set* $Z^m = (Z_1^m, \dots, Z_p^m)$
Next *m*
Return *data sets,* $Z^1, \dots, Z^M$

The resulting reverse-mapped attribute j in the m^{th} synthetic data set Z_j^m expresses Y_j^m as a permutation of X_j. Since the point values of a synthetic attribute are unlikely to be the same as the point values of the original data, particularly in the case of numerical attributes, one must also add E_j^m, the difference between Y_j^m and Z_j^m, to get an exact recomposition of Y_j^m as a function of X_j. Then, and since Z_j^m is a permutation of X_j, it always holds that (with P_j^m denoting a permutation matrix):

$$Y_j^m = \mathrm{P}_j^m X_j + E_j^m, \forall j = 1, \dots, p \text{ and } \forall m = 1, \dots, M \qquad (1)$$

Equation (1) shows that, conceptually, a synthetic data set is functionally equivalent to i) permuting the original data; ii) adding some noise to the permuted data. But, since the noise added has to be necessarily small, as it cannot by construction alter ranks, it does not offer protection of any sort against disclosure risk. In fact, it represents an information loss (as it modifies the marginal distributions of a data set) that is not matched by a decrease in disclosure risk: if, for example, an attacker learns from a data set that the income of an individual is 102 while in reality it is 100, privacy has been violated in the same way as if the intruder was able to retrieve the exact value. Thus, the imprecision due to the small noise is not relevant for privacy. But any anonymization method, synthetic or not, must intuitively comply with the basic principle that any information loss triggered by anonymization must have a counterpart in terms of improved protection. Clearly, the small noise addition does not comply with this principle and can thus be discarded. As a result, the anonymized version of a data set always has an underlying structure that exactly preserves the marginal distributions of the original data (as they are simply a permutation of the original ones), but alters the relative ranks across attributes [15]. Stated otherwise, what ultimately brings protection (and also information loss) are the changes in relationships between attributes.

At first glance, viewing synthetic data as a rank permutation may seem counterintuitive. After all, and as mentioned above, there is no mapping between the synthetic records and the original records. However, the synthetic data set tries to mimic the information in the original data set. In turn, this mimicked information can be expressed as a function of the original data, but with a different rank structure. Thus, at a fundamental level of functioning, a synthesizer can be viewed as a generator of different permutation structures of the original data, or equivalently as a way to generate some permutation matrices for anonymization. The generation of M synthetic data

sets is thus equivalent to the generation of *M* permutation matrices. As it has been previously characterized in the literature that any non-synthetic SDC method is also equivalent to the generation of specific permutation matrices [16,17], *the distinction between synthetic and non-synthetic approaches to anonymization does not seem a fundamental one. As a consequence, synthetic methods must undergo a disclosure risk scrutiny just like their non-synthetic counterparts*.

The ramifications of the above conclusion can further be grasped by recalling the example of a perfect synthesizer. In that case, with a perfect mimic of the information, all multivariate relationships must be exactly preserved. As a result, the permutation matrix has to be the identity matrix (which is a particular case of a permutation matrix where no permutation takes place) and the synthetic data set is the same as the original data set. More realistically, *the better is a synthesizer, the closer to the identity matrix will be each of the underlying permutation patterns contained in the multiple synthetic data sets begin generated.*

Finally, and while the scope of this paper is to investigate the privacy guarantee of synthetic data, it must be noted that the results developed above have broader implications. A releaser could for example decide to release only reverse-mapped synthetic data sets. This solution would not entail additional privacy risks as we saw, but will always offer superior information quality due to the exact preservation of the marginal distributions. Each synthetic data set will thus convey a different rank structure according to the targeted design feature of the original data. Such a possibility is a path for future research.

3.2 Multiple rank-based record linkage attack

The multiple reverse mapping procedure can be easily engineered by the data releaser because he has at his disposal both the original and the synthetic data sets, as in the case of non-synthetic SDC techniques [11]. But as we have argued, in the case of synthetic data, the releaser and the maximum-knowledge attacker are at the same level of knowledge. Thus the attacker, who tries to perform the equivalent of a known-plaintext attack in cryptography, can also reverse map each synthetic data set, eliminate the small noise addition and ultimately be confronted with a collection of data sets that contain only the original data but with different permutation structures. Here, a fundamental departure from non-synthetic anonymization is that the attacker is entitled to several attempts to perform his attack. For instance, if trying to learn say the level of income of an individual, the attacker will try on the *M* data sets to retrieve the value. Intuitively, one can then see that the question of privacy in synthetic data may be trickier than previously thought: the attacker, by retrieving *M* values of income during his attack, could be confused (if the values are very different), comforted (if the values are close), or most likely be helped by narrowing the range of potential values. That is, it is in fact possible that synthetic data may entail a higher degree of privacy risk than non-synthetic anonymized data (in the latter type of data, only one anonymized data set is typically released).

To mount the attack against synthetic data, the recently developed procedure of rank-based record linkage [8] can be repeated *M* times. We consider this specific

linkage type better than other types (such as distance-based linkage or probabilistic linkage), because, as outlined above, data anonymization can be basically described as rank perturbation. Thus, rank-based record linkage appears to be the overarching procedure for evaluating disclosure risk (see [17] for a detailed explanation).

Denote by $O = (o_{ij})$ and $S^m = (s_{ij}^m)$ the rank matrices of the original data set and of the m[th] synthetic data set, respectively[1]. The procedure of multiple rank-based record linkage on synthetic data is as follows:

***Algorithm**: multiple rank-based record linkage*
***Require**: rank matrix O of the original data set*
***Require**: rank matrices $S^1, \dots, S^M$ of the M synthetic data sets $Y^1, \dots, Y^M$*
For** m=1,..., M **do
  For** i=1,...,n **do
    For** l=1,...,n **do
      ***Compute** $d_{il}^m = Criterion[abs(o_{i1} - s_{l1}^m), \dots, abs(o_{ip} - s_{lp}^m)]$*
    ***Next** l*
    Linked index of i in $Y^m = arg\ min_l(d_{il}^m)$
  Next i
***Next** m*
***Return** linked indices of i in the M synthetic data sets*

This procedure is the multi-data set version of the procedure outlined in [8]. It reports the *M* possible matches of an original record with the *M* synthetic data sets. Several criteria can be selected, such as the sum or the minimum of rank differences. To evaluate the privacy guarantees of non-synthetic methods, the criterion will generally depend on the method, e.g. the sum for noise addition or the maximum for data swapping [8]. In the context of synthetic data, this choice is less clear and several criteria should ideally be considered.

4 Empirical illustrations

The objective of this section is to illustrate the concepts of multiple reverse mapping of synthetic data and multiple rank-based record linkage. The experiment is based, without loss of generality, on a small data set of 20 observations and three attributes, and proceeds as follows:

- The assumed original data set is generated by sampling N(50,10^2), N(500,50^2) and N(2500,250^2) distributions, respectively. The correlation coefficient between the first and the second attribute is 0.56, 0.25 between the first and the third, and 0.16 between the second and the third.

[1] Using these notations, o_{ij} is the rank of attribute *j* in original record *i* and s_{lj}^m is the rank of attribute *j* in synthetic record *l* of the m^{th} synthetic data set.

- *M=3* synthetic data sets are generated using a similar sampling procedure. The synthetic data are directly generated with the same size as the original data, albeit one can use the pre-sampling procedure developed above to eventually align the sizes of the former with the size of the latter.
- For the sake of illustration, we consider three different levels of closeness to the original data. As stated previously, the goal of this paper is not to discuss the issue of how to generate a satisfying synthesizer. Rather, by using three different sets, we try to account for the difficulty of generating a satisfying synthesizer:
  - The first synthetic data set is very close to the original data (but does not replicate them perfectly). It was sampled from the same normal distributions from which the original data set was sampled. As a result, the joint relationships between the three attributes are slightly altered (the correlation coefficient between the first and the second synthetic attribute is 0.52, 0.18 between the first and the third and 0.21 between the second and the third).
  - The second synthetic data set has also the joint relationships between the three attributes slightly altered (the correlation coefficient between the first and the second synthetic attribute is 0.44, 0.25 between the first and the third and 0.21 between the second and the third) but with also the properties of the marginal distributions not exactly preserved, i.e. the attributes are sampled from $N(45,8^2)$, $N(450,40^2)$ and $N(2200,200^2)$ distributions, respectively.
  - The third synthetic data set has its marginal distributions sampled from the same as the second one. However, no particular effort is made to preserve the joint relationships (the correlation coefficient between the first and the second synthetic attribute is 0.17, 0.12 between the first and the third and 0.09 between the second and the third).

Table 2 shows the multiple reverse-mapping procedure for the first attribute in the three synthetic data sets[2]. It can be seen that each synthetic data set is expressed as a permutation of the original data. As outlined in the last section, these versions do not entail more disclosure risk than the first generated synthetic data sets, but offer an improved level of information by exactly preserving marginal distributions.

Now, a maximum-knowledge attacker can exactly perform reverse mapping for all attributes and can attempt to recreate the correct linkage. A releaser can also do the same to gauge the privacy of his synthetic data sets before release. Of course, identity disclosure may appear as a peculiar notion for synthetic data but it is still conceivable: an attacker may try to identify which synthetic individuals are the most similar to real individuals, i.e. trying to retrieve some clones. However, we believe that more interesting in the context of synthetic data is attribute disclosure, i.e. when a confidential information contained in the synthetic data sets can be revealed and will closely or exactly correspond to the information of a real individual.

[2] The two other attributes are not shown here due to space constraints but their reverse-mapped versions can be displayed in exactly the same way.

Table 2. Example of multiple reverse mapping on synthetic data sets

Original data set			Synthetic data set 1				Synthetic data set 2				Synthetic data set 3			
ID	X1	Rank of X1	X1	Rank of X1	*Reverse-mapped X1*	Small noises	X1	Rank of X1	*Reverse-mapped X1*	Small noises	X1	Rank of X1	*Reverse-mapped X1*	Small noises
1	38	3	46	9	*51*	-5	33	2	*37*	-4	38	4	*39*	-1
2	66	19	36	1	*31*	5	54	19	*66*	-12	46	14	*57*	-11
3	56	12	43	5	*41*	2	50	16	*63*	-13	42	8	*50*	-8
4	53	11	59	14	*57*	2	37	6	*45*	-8	41	6	*45*	-4
5	31	1	41	4	*39*	2	43	13	*56*	-13	49	16	*63*	-14
6	63	16	61	16	*63*	-2	45	15	*61*	-16	49	17	*63*	-14
7	39	4	44	7	*49*	-5	33	3	*38*	-5	56	20	*70*	-14
8	63	17	56	13	*56*	0	41	11	*53*	-12	42	9	*51*	-9
9	51	9	76	20	*70*	6	40	9	*51*	-11	45	12	*56*	-11
10	56	13	49	10	*51*	-2	37	5	*41*	-4	53	19	*66*	-13
11	70	20	65	17	*63*	2	37	4	*39*	-2	42	7	*49*	-7
12	61	15	59	15	*61*	-2	43	12	*56*	-13	35	3	*38*	-3
13	41	5	40	3	*38*	2	32	1	*31*	1	44	11	*53*	-9
14	49	7	43	6	*45*	-2	51	17	*63*	-12	47	15	*61*	-14
15	51	10	53	12	*56*	-3	58	20	*70*	-12	28	1	*31*	-3
16	64	18	51	11	*53*	-2	39	8	*50*	-11	50	18	*64*	-14
17	45	6	66	18	*64*	2	45	14	*57*	-12	33	2	*37*	-4
18	57	14	44	8	*50*	-6	39	7	*49*	-10	42	10	*51*	-9
19	37	2	72	19	*66*	6	41	10	*51*	-10	40	5	*41*	-1
20	50	8	39	2	*37*	2	53	18	*64*	-11	46	13	*56*	-10

A maximum-knowledge attacker can conduct an attack on a specific attribute by ignoring his knowledge of this attribute in the original data; this is part of the flexibility offered by the maximum-knowledge attacker model (see above and also [2]). The maximum-knowledge attacker can then use the multiple rank-based record linkage procedure to see how well he can recreate the ranks of the ignored attribute; that would simulate a partial-knowledge attacker who did not know the third original attribute and wanted to guess it. Table 3 shows the result of such an attack when knowledge of the third attribute of the original data set is ignored and the sum of rank differences criterion is used to perform multiple rank-based record linkage on the first and second attributes.

Table 3. Example of multiple rank-based record linkage: *third* attribute disclosure scenario

Original data set			Multiple rank-based record linkage: ranks identified by the intruder for X3		
ID	X3	Rank of X3	Synthetic data set 1	Synthetic data set 2	Synthetic data set 3
1	2228	2	11	9	8
2	2299	4	12	18	4
3	2534	10	1	8	12,17
4	2526	9	5	17	11
5	2336	5	16	13	2
6	2598	13	19	19	3
7	2736	16	2	9	8
8	2557	11	12	3	10.9
9	2704	15	17,4,5	16	12,2
10	2513	8	5	17	13
11	2942	19	17	3	10
12	2737	17	18	7	3
13	2559	12	2	2	8
14	2809	18	8	16	16
15	2195	1	4.5	16	11
16	2655	14	6,19	11	4
17	2963	20	15	5	15
18	2298	3	3	7	7
19	2382	6	11	9	8
20	2428	7	15	15	3,14

In this example, one can see that the outcome of an attack on synthetic data can either create confusion to a partial-knowledge attacker, or on the contrary help him narrow his knowledge of the attribute. Consider for example record no. 1 in the original data, with a value of rank 2 for the third attribute. What the attacker gains as information is wrong in each of the synthetic data sets, with a possible rank identified as ranging between 8 and 11. In fact, in that case, having multiple sets consistently orientates the partial-knowledge attacker in the wrong direction. The same is true for several records, e.g. nos. 7, 15, 17. For these individuals, it can be reasonably argued that synthetic data sets offer more privacy in the sense that they fool the attacker consistently across all sets released.

Now consider records nos. 2 and 18. Respectively the third and first synthetic data sets perfectly disclose the attribute values of these records. But because the other sets lead into another direction, the partial-knowledge attacker is again confused. As a result, synthetic data sets seem to provide here again better protection than non-synthetic approaches for these records. However, the partial-knowledge attacker can claim with reasonable confidence that the real value for record no. 2 is between ranks 4 and 18 of the original data and for record no. 18 between 3 and 7. That is, he can claim that the eighteenth individual has a value for the third attribute comprised between 2298 and 2428. Clearly, he has gained some information from the synthetic data sets.

The information can be also narrowed for records where no exact attribute disclosure occurs across the three synthetic data sets in the first place. Consider for example records nos. 4 and 20. For the former, the attacker can claim that the real value is comprised between 2336 and 2737; for the latter, he can claim it is between 2298 and 2704.

Alternatively, assuming that the maximum-knowledge attacker now ignores his knowledge of the first attribute in the original data leads to the similar presence of edges in information (Table 4). For example, for records nos. 9 and 18 the knowledge of the first attribute is narrowed to a significant extent.

While these examples are meant to be illustrative, they however tend to suggest that synthetic data do not come always with low disclosure risk. Releasing multiple data sets can in fact be viewed as an additional privacy threat. Even if by definition no real individual is present in the synthetic data, some clones nonetheless are, and these clones can be re-identified to learn some information about some real individuals.

Originally, the proposal of releasing multiple data sets aimed at enhancing the quality of information offered by synthetic data. But, considering that such practice can be undoubtedly cumbersome for the users and that the quality of information can in some cases be made at least as good with a single data set [10], having multiple releases seem also to entail some previously uncharacterized privacy risks that render this practice questionable.

Table 4. Example of multiple rank-based record linkage: *first* attribute disclosure scenario

Original data set			Multiple rank-based record linkage: ranks identified by the intruder for X1		
ID	X1	Rank of X1	Synthetic data set 1	Synthetic data set 2	Synthetic data set 3
1	38	3	15,1	6	14
2	66	19	12	16,17	20
3	56	12	20	2,4	19,15
4	53	11	3,6	19	7,10
5	31	1	12,11	13	20
6	63	16	7	19,5	4,11
7	39	4	1,7,18	10	17
8	63	17	19	2,4	9
9	51	9	16	8	12
10	56	13	6	1,14	7,3
11	70	20	10	20	5
12	61	15	18	12,10	6
13	41	5	8,2	3	17,1
14	49	7	17	8	8
15	51	10	12	17	2
16	64	18	16	5	4
17	45	6	1	15	18
18	57	14	15	7	16
19	37	2	9	7	1,13
20	50	8	9,3,14	1,14	7,3,13

5 Conclusions and future work

Synthetic data is often perceived as having lower disclosure risk than other forms of SDC methods. In this paper, we show that this may not always be the case. Despite the fact that no real individuals are included in a data release, at least as far as fully synthetic data are concerned, synthetic and real individuals remain however linked by the information they convey. If an attacker is able to retrieve some information on real individuals that happens to be correct, it ultimately does not matter that this information is based on simulated data. Even if such a disclosure does not fall under the consideration of any legislation on privacy, it can nonetheless be viewed as unethical insofar as it affects real individuals.

The objective of this paper was thus to investigate the privacy guarantee of synthetic data. Using recent advances in the literature on the definition of an attacker in data anonymization, we confronted synthetic data to an attack by a maximum-knowledge intruder. While conservative in its stance, this model has the merit to establish a common benchmark to gauge the privacy guarantees of non-synthetic anonymization methods. It thus seems only fair to consider synthetic data in the same context. Actually, the maximum-knowledge attacker is the counterpart of the popular and widely used notion of known-plaintext attack in cryptography.

We first presented an extension of a reverse-mapping procedure that can be performed both by an attacker and a synthetic data releaser. Under a reasonable assumption on the size of the synthetic data sets to be released, this procedure shows that in fact any synthetic data set can always be expressed as a permutation of the original data, in a way similar to non-synthetic SDC techniques. This result offers applications beyond disclosure risk assessment. For one thing, it is always possible to release synthetic data sets with the same privacy properties but with an improved level of infor-

mation, because the marginal distributions can be always preserved without increasing risk. On the privacy front, reverse mapping leads to the consequence that the distinction made in the literature between non-synthetic and synthetic data is not so clear-cut. Thus, both approaches must be evaluated against the same privacy challenges.

Next, we proposed an extension of the rank-based record linkage procedure that can also be performed both by the attacker and the synthetic data releaser. In particular, the latter can use it to assess the privacy guarantee of its synthetic data before release. This procedure shows that the practice of releasing several synthetic data sets for a single original data set entails privacy issues that do not arise in non-synthetic anonymization (where typically only one anonymized data set is released). Indeed, the multiple releases can lead to better privacy guarantees, by confusing the attacker, or facilitate attribute disclosure by helping the attacker narrow the range of the possible values that he is trying to retrieve. An empirical investigation in the last section illustrates those issues. We believe that this has interesting consequences for synthetic data releases that deserve further investigation.

The results presented in this paper are preliminary and illustrative. As future work, we plan to: i) investigate the theoretical and empirical conditions under which multiple synthetic data sets can lead to more confusion than help for the attacker; ii) assess the possibility of considering synthesizers as tools to generate different permutation patterns, which could offer some insights for non-synthetic anonymization techniques; iii) enlarge the scope of the experimental work by using various synthetic data sets, and in particular assess the occurrence and the magnitude of range narrowing during an attack.

Acknowledgments and disclaimer

The following funding sources are gratefully acknowledged by the third author: European Commission (project H2020-700540 “CANVAS”), Government of Catalonia (ICREA Acadèmia Prize) and Spanish Government (projects TIN2014-57364-C2-1-R “SmartGlacis” and TIN2015-70054-REDC). The views in this paper are the authors’ own and do not necessarily reflect the views of UNESCO or any of the funders.

References

1. J. Domingo-Ferrer and K. Muralidhar, "New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users", *Information Sciences*, Vol. 337, pp. 11-24, Apr 2016.
2. J. Domingo-Ferrer, S. Ricci and J. Soria-Comas, "Disclosure risk assessment via record linkage by a maximum-knowledge attacker", *13th Annual International Conference on Privacy, Security and Trust-PST 2015*, Izmir, Turkey, Sep 2015.
3. J. Domingo-Ferrer, D. Sánchez and G. Rufian-Torrell, "Anonymization of nominal data based on semantic marginality", *Information Sciences*, Vol. 242, pp. 35-48, May 2013.
4. J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control*, Springer, 2011.

5. J. Drechsler, S. Bender and S. Rässler, "Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel", *Transactions on Data privacy*, Vol. 1, pp. 105-130, 2008.
6. J. Hu, J. P. Reiter and Q. Wang, "Disclosure risk evaluation for fully synthetic data", *Lecture Notes in Computer Science, Vol. 8744 (Privacy in Statistical Databases - PSD2014)*, pp. 185-199, Sep 2016.
7. A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer and P.-P. de Wolf, *Statistical Disclosure Control*, Wiley, 2012.
8. K. Muralidhar and J. Domingo-Ferrer, "Rank-based record linkage for re-identification risk assessment", *Lecture Notes in Computer Science, Vol. 9867 (Privacy in Statistical Databases - PSD2016)*, pp. 225-236, Sep 2016.
9. K. Muralidhar and J. Domingo-Ferrer, "Microdata Masking as Permutation," *UNECE/EUROSTAT Work Session on Statistical Data Confidentiality*, Helsinki, Finland, October 2015.
10. K. Muralidhar and R. Sarathy, " A Comparison of Multiple Imputation and Data perturbation for Masking Numerical Variables ", *Journal of Official Statistics*, Vol. 22, pp. 507-524, 2006.
11. K. Muralidhar, R. Sarathy and J. Domingo-Ferrer, "Reverse mapping to preserve the marginal distributions of attributes in masked microdata", Lecture Notes in Computer Science, Vol. 8744 (Privacy in Statistical Databases - PSD 2014), pp. 105-116, Sep 2014.
12. J. P. Reiter, Q. Wang and B. Zhang, "Bayesian estimation of disclosure risks in multiply imputed, synthetic data", *Journal of Privacy and Confidentiality*, Vol. 6:1, Article 2, 2014.
13. J. P. Reiter, "Satisfying disclosure restrictions with synthetic data sets", *Journal of Official Statistics*, Vol. 18, pp. 531-544, 2002.
14. J. P. Reiter, "Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study", *Journal of The Royal Statistical Society Series A*, Vol. 168, pp. 185-205, 2005.
15. D. B. Rubin, "Discussion: statistical disclosure control limitation", *Journal of Official Statistics*, Vol. 9, pp. 462-468, 1993.
16. N. Ruiz, "On some consequences of the permutation paradigm for data anonymization: Centrality of permutation matrices, universal measures of disclosure risk and information loss, evaluation by dominance", *Information Sciences*, Vol. 430–431, pp. 620-633, March 2018.
17. N. Ruiz, "A General cipher for individual data anonymization", *under review for Information Sciences* (https://arxiv.org/abs/1712.02557), 2017.
18. J. Soria-Comas and J. Domingo-Ferrer, "A non-parametric model for accurate and provably private synthetic data sets", *Proc. of International Conference on Availability, Reliability and Security-ARES 2017*, art. no. 3. ACM, 2017.
19. L. Willenborg and T. De Waal, *Elements of Statistical Disclosure Control*, Springer, 2001.