Should probably talk about this too :
A Baseline for Attribute Disclosure Risk in Synthetic Data

A Closer Look at the CAP Risk Measure for Synthetic Datasets

Mathieu Baillargeon and Anne-Sophie Charest

Université Laval anne-sophie.charest@mat.ulaval.ca

Abstract. While synthetic datasets have gained in popularity as a statistical disclosure limitation method, there is no clear consensus as to how to measure the privacy protection offered by these datasets. One proposal is the CAP statistic, presented in [1]. In this paper, we take a closer look at the characteristics of this risk measure. In particular, we show how to construct a synthetic dataset which will minimize the CAP value. We also empirically look at the relationship between the average CAP value, which is suggested as the risk measure, and the individual CAP values for each of the individuals in the dataset.

1 Introduction

The use of synthetic datasets for privacy protection has gained in popularity over the years. Still, it is widely accepted that they do not necessarily preserve the privacy of the respondents. Indeed, while synthetic data protect against re-identification disclosure, inferential disclosure is still a concern and so the privacy protection of such synthetic datasets must be measured carefully.

In some cases, the process by which the datasets are generated satisfies differential privacy (as in [2],[3],[4]), and so the privacy guarantees are clear and known. In other cases, one must measure the privacy protection on the synthetic dataset itself. A few measures of risk have been proposed to do so, for e.g. in [5], [6] and [7]. In this paper, we take a closer look at the CAP risk measure, proposed in [1] and highlighted in [8]. Although [9], the forerunner to [1], provides an idea for a generalization for continuous variables, we restrict ourselves to the version for categorical datasets.

We first consider the individual CAP privacy measure in section 2, proving a simple result concerning the relationship between the CAP values for individuals with identical keys and different targets. In section 3, we present the two variants of the average CAP measure for a dataset, and show how to generate synthetic datasets to minimize these measures. Section 4 gives some empirical results regarding the relationship between the individual and the average CAP values. We finally discuss the results in section 5.

2 Individual CAP Measure and Results

We present here a few results on the CAP statistic, as presented in [1], taken as measure of risk for each individual.

2.1 Notation and terminology

Let O denote a dataset, and x_ℓ the ℓ-th of the n records of O. We consider a specific sensitive variable whose values must be protected. All of the possible values for this variable are called the *targets* and are denoted $T_1, \ldots, T_I$.

An intruder would try to predict the value of a target using some or all of the other variables contained in O. Each possible combination of the values of these variables is referred to as a *key*. The possible keys are denoted $K_1, \ldots, K_J$. Note that it is possible that some of those combinations are not present in O.

Thus, each record x_ℓ is associated with a single key $K(x_\ell)$ and a single target $T(x_\ell)$. It is convenient to express the information about the keys and targets of observations in dataset O as a contingency table, where

$$o_{ij} = \sum_{\ell=1}^{n} \mathbb{1}\big[T(x_\ell) = T_i, K(x_\ell) = K_j\big]$$

with $\mathbb{1}$ denoting the indicator function.

2.2 Definition and interpretation

Consider another (non-empty) dataset S which contains the same keys and targets as an original dataset O. Let y_ℓ be the ℓ-th of the m records of S. We can also present S as a contingency table, where $s_{ij} = \sum_{\ell=1}^{m} \mathbb{1}\big[T(y_\ell) = T_i, K(y_\ell) = K_j\big]$. This dataset, which we will refer to as a synthetic dataset, is a candidate to replace O in order to protect the confidentiality of the original respondents. It will often be the case that $m = n$, but it is not required.

Definition 1 (Individual CAP). *The Correct Attribution Probability (CAP) of record x_0 in O with synthetic dataset S is given as*

$$CAP_{x_0}(S) = \begin{cases} \frac{\sum_{\ell=1}^{m} \mathbb{1}[T(y_\ell)=T(x_0)\ ,\ K(y_\ell)=K(x_0)]}{\sum_{\ell=1}^{m} \mathbb{1}[K(y_\ell)=K(x_0)]} & \text{if } \sum_{\ell=1}^{m} \mathbb{1}[K(y_\ell) = K(x_0)] \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The $\mathrm{CAP}_{x_0}(S)$ is thus the proportion of the records in S that have the same target as x_0 among the records that have the same key as x_0 in S. The intuition behind this is that an intruder, knowing the key of the individual x_0, will try to predict the value of $T(x_0)$ based on the distribution of the targets for the records having key $K(x_0)$ in the published dataset S. Smaller values for $\mathrm{CAP}_{x_0}(S)$ are thus preferred.

Note that our definition of $\mathrm{CAP}_{x_0}(S)$ is slightly different from the one in [1] because we treat directly the case where there are no observations with the same key as x_0 in the synthetic dataset instead of leaving them as undefined until we average the CAP values over observations in O.

Note also that we can compute the CAP of record x_0 with O, i.e. the CAP if the original dataset is published. In that case, the denominator will never be zero, so it is given by $\mathrm{CAP}_{x_0}(O) = \frac{\sum_{\ell=1}^{n} \mathbb{1}[T(x_\ell)=T(x_0)\ ,\ K(x_\ell)=K(x_0)]}{\sum_{\ell=1}^{n} \mathbb{1}[K(x_\ell)=K(x_0)]}$. An individual with $\mathrm{CAP}_{x_0}(O) > \mathrm{CAP}_{x_0}(S)$ is deemed to have been been protected by publishing dataset S instead of dataset O.

2.3 Illustration

To help understand the measure, we illustrate it on a simple example. Table 1 shows a dataset with two variables : the smoking status of a person and her health status. It is plausible that an intruder would know the smoking status of an individual and want to use O to learn about her health status, so we set the health status as the target variable. In order to protect the personal information in O, the synthetic dataset S, given in table 2 is considered for publication.

	Smoking	Non-smoking
Sick	20	5
Healthy	30	45

Table 1: Dataset O

	Smoking	Non-smoking
Sick	8	7
Healthy	25	60

Table 2: Dataset S

Consider x_0, a sick smoking person. Then, $\text{CAP}_{x_0}(O) = 20/50 = 0.4$ and $\text{CAP}_{x_0}(S) = 8/33 = 0.2424$. Theoretically, the CAP measure indicates that individual x_0 is better protected if S is published than O since it provides less information about the value of $T(x_0)$.

One may note however that in practice, it does not appear that $T(x_0)$ is really less protected with one of the datasets. Indeed, in both cases, an intruder would likely predict that x_0 is healthy, thus making a wrong prediction. This is thus a first limitation of the CAP measure of risk, which is discussed in more details in [10].

Back to the illustration, consider now x_1 is a healthy smoking person. Then $\text{CAP}_{x_1}(O) = 30/50 = 0.6$ and $\text{CAP}_{x_1}(S) = 25/33 = 0.7576$. Here the CAP measure indicates that individual x_1 would be better protected if O were published instead S. But again, this is not the whole story. In fact, both S and O are bad for x_1 since an intruder would likely predict the correct value for $T(x_1)$, no matter which dataset she has access to.

Considering simply the individual CAP measure, and ignoring the prediction issue for now, we can note that if S were to be released instead of O, the situation would get better for x_0, but worse for x_1. In fact, one may prove a simple result which shows that the CAP values for individuals with identical keys are always related in a similar fashion.

2.4 Result

First, note that there are only a certain number of possible individual CAP values for individuals in a dataset O since all individuals with the same key and target have the same CAP. Now, we can show that the sum of these CAP values over all cells with an identical key and different targets always equals 1.

Theorem 1. *Let K_0 be a certain key and $\{T_1, ..., T_I\}$ be the set of all the possible targets in O. If there is at least one record associated with K_0 in S, then*

$$\sum_{i=1}^{I} \frac{\sum_{\ell=1}^{m} \mathbb{1}[T(y_\ell) = T_i \ , K(y_\ell) = K_0]}{\sum_{\ell=1}^{m} \mathbb{1}[K(y_\ell) = K_0]} = 1.$$

Proof.

$$\sum_{i=1}^{I} \frac{\sum_{\ell=1}^{m} \mathbb{1}[T(y_\ell) = T_i \ , \ K(y_\ell) = K_0]}{\sum_{\ell=1}^{m} \mathbb{1}[K(y_\ell) = K_0]} = \frac{\sum_{\ell=1}^{m} \left(\sum_{i=1}^{I} \mathbb{1}[T(y_\ell) = T_i \ , \ K(y_\ell) = K_0] \right)}{\sum_{\ell=1}^{m} \mathbb{1}[K(y_\ell) = K_0]}$$

$$= \frac{\sum_{\ell=1}^{m} \mathbb{1}[K(y_\ell) = K_0]}{\sum_{\ell=1}^{m} \mathbb{1}[K(y_\ell) = K_0]} = 1.$$

■

Theorem 1 tells us that improving the CAP of individuals in a certain cell will generally worsen that of individuals in a different cell with the same key. A way to avoid this effect is by ensuring that the key K_0 does not appear at all in the synthetic dataset S. We will see in the following theorems what impact this has on the optimization of the average of the individual CAP measures.

3 Average CAP Measures and Results

3.1 Definitions

In practice [1] suggests to look at the average of the CAP individual measures in order to identify the most protective synthetic dataset among a group of candidates. In fact this is what is referred to as the CAP for dataset O. There are two ways to define this average CAP, which differ by the treatment of the observations whose key does not appear in the synthetic dataset.

Definition 2 (Version 1 of average CAP). *Let O be an original datasets with observations $x_1, \ldots, x_n$ and S a synthetic dataset for O. The average CAP of O with S is given by*

$$\overline{CAP}_O(S) = \frac{1}{n} \sum_{\ell=1}^{n} CAP_{x_\ell}(S).$$

This is simply the average of the individual CAP as defined in Definition 1. Remember that the value of $\text{CAP}_S(x_\ell)$ will be 0 if $K(x_\ell)$ does not appear in S. An alternative, proposed in [1] is to ignore these particular records when calculating the mean CAP. This leads to the second version of the CAP statistic:

Definition 3 (Version 2 of average CAP). *Let O be an original datasets with observations $x_1, \ldots, x_n$ and S a synthetic dataset for O. Define the set $A = \{x_\ell \in O \mid \nexists\, y_p \in S \text{ s.t. } K(x_\ell) = K(y_p)\}$. In this case, the average CAP of O with S is given by*

$$\overline{CAP}^*_O(S) = \frac{1}{n - |A|} \sum_{\ell=1}^{n} CAP_{x_\ell}(S).$$

Note that if $|A| = n$, for example if there are no observations associated with a certain key in O and all synthetic observations are associated with this key, then $\overline{\text{CAP}}^*_O(S)$ will be undefined. We will ignore this case in the sequel.

3.2 Minimizing $\overline{\text{CAP}}$ and $\overline{\text{CAP}}^*$

Both versions of the average CAP measure are measures of risk, which one would thus want to minimize. Previous work with CAP has empirically compared the values of CAP for different datasets. Here, we show how one can, for any original data O, produce a synthetic dataset S which minimizes the average CAP measures. The results are slightly different for the two versions, but we first prove a result which is valid for both. It shows how to minimize the average CAP among synthetic datasets with a given set of marginal totals for the keys.

We first need a little more notation. Define $o^j_{min} = \min_{i=1,\ldots,I} o_{ij}$, the minimum cell value among those with key K_j, and $o_{min} = \min_{j=1,\ldots,J}(o^j_{min})$ the overall minimum cell value. Define also the three sets $B = \{(i,j) \mid o_{ij} = o_{min}\}$, $C_j = \{(i,j) \mid o_{ij} = o^j_{min}\}$ and $C = \bigcup_{j=1}^{J} C_j$.

Lemma 1. *Fix an original dataset O. Consider $\mathbb{S}$ the set of all possible non-empty synthetic datasets for O of size m and with marginal totals $(m_1, \ldots, m_J)$ for the keys. Then, a dataset $S \in \mathbb{S}$ has the smallest $\overline{CAP}_O$ of all datasets in $\mathbb{S}$ if and only if all its cells are empty, except possibly those with coordinates in C.*

Proof. We prove the lemma using $\overline{CAP}_O$. The proof is identical for $\overline{CAP}^*_O$.

Note that because we fix the marginal totals $(m_1, \ldots, m_J)$ over keys, all we can do to minimize the average CAP is modify the target value of some of the synthetic observations, for one or several of the keys.

Since we consider non-empty datasets, there is at least one key K_α such that $m_\alpha \neq 0$. Consider two datasets $S_0, S_1 \in \mathbb{S}$ which are identical except that an individual associated with key K_α has been moved from target T_β (in S_0) to target T_η (in S_1). Let s_{ij} denote the cell counts for target T_i and key K_j in the dataset S_0. The difference between the average CAP for the two datasets is

$$
\begin{aligned}
&\overline{\text{CAP}}_O(S_1) - \overline{\text{CAP}}_O(S_0) \\
&= \frac{1}{n}\left(o_{11}(0) + \ldots + o_{\beta\alpha}\left(\frac{s_{\beta\alpha}-1}{m_\alpha} - \frac{s_{\beta\alpha}}{m_\alpha}\right) + \ldots + o_{\eta\alpha}\left(\frac{s_{\eta\alpha}+1}{m_\alpha} - \frac{s_{\eta\alpha}}{m_\alpha}\right) + \ldots + o_{IJ}(0)\right) \\
&= \frac{1}{n}\left(\frac{o_{\eta\alpha} - o_{\beta\alpha}}{m_\alpha}\right).
\end{aligned}
$$

If $o_{\eta\alpha} < o_{\beta\alpha}$, then $\overline{\text{CAP}}_O(S_1) < \overline{\text{CAP}}_O(S_0)$, meaning that S_1 has a better average CAP than S_0. One can repeat the same process to continue improving the average CAP until it is impossible to find targets T_β and T_η such that $o_{\eta\alpha} < o_{\beta\alpha}$. Note that if $o_{\eta\alpha} = o_{\beta\alpha}$, then $\overline{\text{CAP}}_O(S_1) = \overline{\text{CAP}}_O(S_0)$ which means that the change has no effect on the average CAP. Hence, for a given key K_α the minimum CAP will be attained by a dataset with all synthetic observations in cells in C_j.

Repeat the argument with all other non-empty keys to get the final result. ∎

Lemma 1 shows how to minimize the CAP measure for a given distribution of the synthetic observations among the keys. To obtain the overall minimal average CAP statistic we must now consider all possible distributions over the keys. The results now depend on the version of the average CAP measure used. We consider each case at a turn.

For version 1, we find that one way to minimize the average CAP is by placing all synthetic observations in a single cell, chosen among all cells in O with the minimal number of individuals.

Theorem 2 (Minimal $\overline{CAP}_O$ and best synthetic dataset). *Fix an original dataset O. For any size of synthetic dataset $m > 0$, the smallest possible value of $\overline{CAP}_O$ is o_{min}/n and any synthetic dataset S such that every cell is empty except one, whose coordinates are in B, reaches this value.*

Proof. First note that we only consider non-empty synthetic datasets. Otherwise, the smallest possible value for $\overline{CAP}_O$ would simply be zero.

Fix m, the number of records in the synthetic dataset. Lemma 1 provides an exhaustive list of candidate synthetic datasets for each set of marginal counts on keys $m_1, \ldots, m_J$ such that $\sum_{j=1}^{J} m_j = m$. All candidate synthetic datasets for a given set of $\{m_i\}_{i=1}^{J}$ have the same value of $\overline{CAP}_O$, so without loss of generality we only consider one candidate dataset for each set of $\{m_i\}_{i=1}^{J}$, namely a dataset with all empty cells, except for a single one for each key, with coordinates in C.

Now, note that of all these datasets, the ones that have the same empty keys have the same $\overline{CAP}_O$. Indeed, consider two dataset S_1 and S_2 with the same set of empty keys, and S_2, built such that r individuals in S_1 were moved from K_α to other non-empty keys (and possibly other targets, depending on their new keys) without K_α becoming empty. Then,

$$\begin{aligned}\overline{CAP}_O(S_1) - \overline{CAP}_O(S_2) &= \frac{1}{n}\left(o^1_{min}(0) + \ldots + o^\alpha_{min}(0) + \ldots + o^J_{min}(0)\right)\\ &= 0.\end{aligned}$$

Here, each o^j_{min} is in fact multiplied by a factor of $(1-1)$ or $(0-0)$, depending on if K_j is empty or not in S_1 and S_2.

One may however reduce the value of the average CAP by creating new empty keys. Indeed, consider S_1 again, as well as S_3, a dataset built from S_1 by moving all records associated with K_α to other non-empty keys of S_1 (and possibly other targets, depending on the new keys). This means that S_3 has the same empty keys than S_1 in addition to K_α. Then,

$$\begin{aligned}\overline{CAP}_O(S_1) - \overline{CAP}_O(S_3) &= \frac{1}{n}\left(o^1_{min}(0) + \ldots + o^\alpha_{min}(1-0) + \ldots + o^J_{min}(0)\right)\\ &= \frac{o^\alpha_{min}}{n} > 0\end{aligned}$$

showing that S_3 has a smaller $\overline{CAP}_O$ than S_1. We can apply the same argument until all observations in S_1 are associated to a single non-empty key. En fait,

rien ne dit que $o^{\alpha}_{min} > 0$. Dans le cas ou cette quantité est égale à zéro, on a aucun changement. On doit donc spécifier qu'en cas d'égalité, on conserve le jeu avec le plus de clés vides possible.

The synthetic dataset with minimal $\overline{CAP}_O$ is thus one where all m individuals have the same key and the same target. For any such dataset S, $\overline{CAP}_O(S) = o^{\beta}_{min}/n$ where K_β is the only non-empty key. This will be minimal and equal o_{min}/n as long as all m synthetic observations are placed in a cell with coordinates in B, which proves the result. ■

It should not come as a surprise that minimizing the $\overline{CAP}_O$ yields a pretty useless dataset : no risk usually implies little or no information. It is interesting however to be able to calculate simply the minimal $\overline{CAP}_O$ measure for any dataset. Note that this value does not depend on the size of the synthetic dataset, only on the combinations of the keys and targets in the original dataset. Also, the form of a minimal synthetic datasets gives us some information about how the risk measure works, namely that an easy way to reduce it is by simply assigning zero values to cells with a large number of individuals in the original dataset.

We find a similar result with $\overline{CAP}^*_O$.

Theorem 3. *Fix an original dataset O. Define o^*_{min} as the smallest possible value of $o_{ij}/o_{\bullet j}$ in O and define the set $D = \{(i,j)|o_{ij}/o_{\bullet j} = o^*_{min}\}$, where $o_{\bullet j} = \sum_{i=1}^{I} o_{ij}$, following usual notation. The smallest possible value for $\overline{CAP}^*_O$ is o^*_{min} and a dataset S such as every cell is empty except one, whose coordinates are in D, reaches this value.*

Proof. We will prove this by contradiction.
Imagine there is a dataset $\tilde{S}$ such that $\overline{CAP}^*_O(\tilde{S}) < o^*_{min}$ and it is minimal among all possible synthetic datasets for O. In particular, $\tilde{S}$ may contain some empty keys. Define $N = \{j \mid \tilde{s}_{\bullet j} \neq 0\}$ and $n^* = \sum_{j \in N}(o_{\bullet j})$. Since lemma 1 is also valid for $\overline{CAP}^*_O$, we know that all the coordinates of the non-empty cells of $\tilde{S}$ must be in C.

If $\tilde{S}$ only has only one non-empty key K_β, then $\overline{CAP}^*_O(\tilde{S}) = o^{\beta}_{min}/o_{\bullet\beta}$. But then we would have $\overline{CAP}^*_O(\tilde{S}) \geq o^*_{min}$, contradicting the fact that $\overline{CAP}^*_O(\tilde{S}) < o^*_{min}$. So, $\tilde{S}$ must have more than one non-empty key.

Also, if $\frac{o^{\beta}_{min}}{o_{\bullet\beta}} = \frac{o^{j}_{min}}{o_{\bullet j}}\ \forall j \in N$, then $\mathrm{CAP}(\tilde{S}) = \frac{\sum_{j\in N} o^{j}_{min}}{\sum_{j\in N} o_{\bullet j}} = \frac{o^{\beta}_{min}}{o_{\bullet\beta}} \geq o^*_{min}$, contradicting again the fact that $\overline{CAP}^*_O(\tilde{S}) < o^*_{min}$. So, there must be K_α, possibly with $\alpha = \beta$, such that $\frac{o^{\alpha}_{min}}{o_{\bullet\alpha}} \geq \frac{o^{j}_{min}}{o_{\bullet j}}\ \ \forall j \in N$ and $j_0 \in L$ such that $\frac{o^{\alpha}_{min}}{o_{\bullet\alpha}} > \frac{o^{j_0}_{min}}{o_{\bullet j_0}}$,where $L = N \setminus \alpha$. In particular, it means that

$$\sum_{j\in N}(o_{\bullet j} \times o^{\alpha}_{min} - o_{\bullet\alpha} \times o^{j}_{min}) > 0 \tag{1}$$

since the term for j_0 will be positive. We will need this result soon.

For now, consider $\tilde{S}_2$, a dataset built from $\tilde{S}$ by moving all the individuals associated with K_α to other non-empty keys of $\tilde{S}$ (and possibly other targets, depending on their new keys). This means that $\tilde{S}_2$ has one more empty key than $\tilde{S}$. Then,

$$\overline{\text{CAP}}^*_O(\tilde{S}) = \frac{\sum_{j\in N}(o^j_{min})}{n^*} \qquad \text{and} \qquad \overline{\text{CAP}}^*_O(\tilde{S}_2) = \frac{\sum_{j\in L}(o^j_{min})}{n^* - o_{\bullet\alpha}}.$$

Combining equation (1) with the fact that $\frac{o^\alpha_{min}}{o_{\bullet\alpha}} \geq \frac{o^j_{min}}{o_{\bullet j}} \quad \forall j \in L$, we find that

$$\sum_{j\in N}(o_{\bullet j} \times o^\alpha_{min} - o_{\bullet\alpha} \times o^j_{min}). > 0$$

Cette dernière phrase semble redondante. Je crois que je la retirerais.
A little algebra now gives the desired result:

$$\begin{aligned}
&\sum_{j\in N}(o_{\bullet j} \times o^\alpha_{min} - o_{\bullet\alpha} \times o^j_{min}) > 0 \\
\Rightarrow &\Big(\sum_{j\in N} o_{\bullet j}\Big) o^\alpha_{min} - o_{\bullet\alpha}\Big(\sum_{j\in N} o^j_{min}\Big) > 0 \\
\Rightarrow &\; n^*(o^\alpha_{min}) - o_{\bullet\alpha}\Big(\sum_{j\in N} o^j_{min}\Big) > 0 \\
\Rightarrow &\; (n^* - o_{\bullet\alpha})\Big(\sum_{j\in N} o^j_{min}\Big) - n^*\Big(\sum_{j\in L} o^j_{min}\Big) > 0 \\
\Rightarrow &\; \frac{\sum_{j\in N}(o^j_{min})}{n^*} - \frac{\sum_{j\in L}(o^j_{min})}{n^* - o_{\bullet\alpha}} > 0 \\
\Rightarrow &\; \overline{\text{CAP}}^*_O(\tilde{S}) > \overline{\text{CAP}}^*_O(\tilde{S}_2).
\end{aligned}$$

This is a contradiction since we set $\overline{\text{CAP}}^*_O(\tilde{S})$ to be minimal, proving the theorem. ■

4 Relationship between individual and average CAP

Having studied in detail both the individual CAP values and the two proposed averages for these values, we now turn to study the relationship between the individual CAP measures and these averages, to see to what point minimizing the average CAP adequately protects each of the individuals of the dataset. Here, with proceed empirically.

We consider 4 different original datasets of 900 individuals, given in Tables 3 to 6. O_1 is such that the records are evenly distributed between each cells, so that there is no information in the keys about the values of the targets. At the

opposite, O_2 is constructed so that each key completely determines the target of the individual. The other two datasets are somewhere in between : in O_3 we consider a dataset where every record associated to K_1 is also associated to the target T_3 and O_4 is such that every record associated with T_1 is also associated with K_3.

	K_1	K_2	K_3
T_1	100	100	100
T_2	100	100	100
T_3	100	100	100

Table 3: Original dataset O_1

	K_1	K_2	K_3
T_1	400	0	0
T_2	0	250	0
T_3	0	0	350

Table 4: Original dataset O_2

	K_1	K_2	K_3
T_1	0	176	204
T_2	0	78	93
T_3	127	163	59

Table 5: Original dataset O_3

	K_1	K_2	K_3
T_1	0	0	62
T_2	197	136	134
T_3	74	99	198

Table 6: Original dataset O_4

We randomly select 20 among all the possible possible 3×3 datasets containing 900 records to be used as synthetic datasets. Figure 1 shows the distributions of the invididual CAP values for each of these synthetic datasets. The boxplots are ordered according to version 1 of $\overline{\text{CAP}}_{O_i}$, shown as red dots over each of the boxplots, which is why the order of the boxplots differ for the different original datasets. In each panel, the horizontal blue line shows the average CAP of the original dataset. Note that in fact, there are only 9 possible values for the individual CAP measure for each synthetic dataset since there are only 9 cells, so that each boxplot is created from at most nine values.

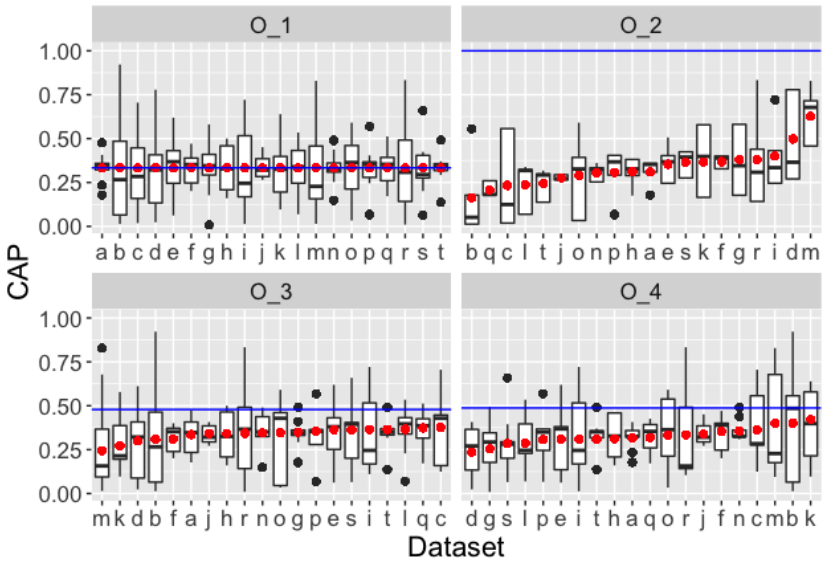

Fig. 1: Distribution of the individual CAP for each synthetic dataset

Figure 1 shows that the distributions of the individual CAP values vary a lot across the synthetic datasets. In particular, datasets with similar or identical average CAP can have completely different individual CAP distributions, hence protecting very differently the original records. In some cases, the average CAP does not seem to summarize very well the risk of the individuals. In particular, the datasets with the smallest average CAP may still be associated with some high individual CAP values, meaning that some of the cells are still at risk. For example, for O_3 dataset b has one of the lowest $\overline{CAP_{O_3}}$ but also the highest individual CAP of all sampled synthetic datasets. These results suggest that one should examine the individual CAP values when selecting a synthetic dataset, and not merely consider the average CAP.

On the same note, consider the synthetic datasets b and g, which we will note S_b and S_g, given in Tables 7 and 8.

	K_1	K_2	K_3
T_1	24	36	15
T_2	226	113	213
T_3	216	54	3

Table 7: Synthetic dataset S_b

	K_1	K_2	K_3
T_1	129	137	99
T_2	2	131	147
T_3	91	111	53

Table 8: Synthetic dataset S_g

As replacement for O_3, S_b would be better than S_g if we were to examine only the average CAP : $\overline{CAP}_{O_3}(S_b) \approx 0.31$ while $\overline{CAP}_{O_3}(S_g) \approx 0.35$. However, we can see on Figure 1 that there are many more individuals with poor protection, that is larger individual CAP values, with dataset S_b.

One may also think about the protection offered by synthetic datasets in terms of the predictive success, using the most common value for each key in the released dataset as discussed previously. Consider for example synthetic datasets a and e, given in Tables 5 and 6, as replacements for dataset O_4.

	K_1	K_2	K_3
T_1	114	118	77
T_2	104	44	134
T_3	105	86	118

Table 9: Synthetic dataset S_a

	K_1	K_2	K_3
T_1	67	97	153
T_2	37	114	101
T_3	169	14	148

Table 10: Synthetic dataset S_e

If we were to examine only the average CAP, S_e would be better than S_a since $\overline{CAP}_{O_4}(S_e) \approx 0.309$ while $\overline{CAP}_{O_4}(S_a) \approx 0.318$. However, only 134 records would be correctly predicted with S_a in case of an attack, while 272 would be with S_e. This shows again that the CAP measures only provides partial information about inferential disclosure.

Another important observation from Figure 1 is that in all of the synthetic datasets the average CAP of the randomly selected datasets are smaller or equal to the average CAP of the original dataset. This is especially surprising since

[1] suggests that a way to evaluate if a synthetic dataset S is safe enough to be released to replace O or not is to compare $\text{CAP}_O(S)$ to $\text{CAP}_O(O)$. It would seem that this condition is much too easy to satisfy, which we now verify empirically by sampling more synthetic datasets.

Figure 2 shows the distribution of the two versions of the average CAP of 2000 randomly selected datasets considered to replace each of O_1, O_2, O_3 or O_4. Again, the blue line represents $\overline{CAP}_{O_i}(O_i) = \overline{CAP}^*_{O_i}(O_i)$, $i = 1, ..., 4$. In all cases, we find that the vast majority of these synthetic datasets indeed have a smaller CAP average than the original dataset.

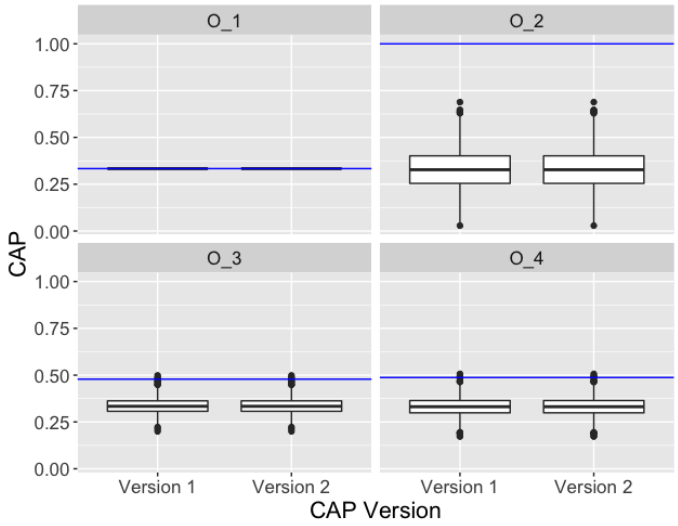

Fig. 2: Distribution of the average CAP of a random sample of 2000 among all the possible synthetic datasets for each of the 4 original datasets O_1, O_2, O_3 and O_4

To understand how almost all synthetic datasets can be better than the original, it is useful to go back to Lemma 1. As we saw in the proof, a way to reduce the value of the average CAP of a synthetic dataset S replacing O is to move people of a certain key from a cell s_{ij} such that o_{ij} is big to a cell s_{kl} such that o_{kl} is small. Of course, the converse is also true. Now, notice that $\overline{CAP}_O(O)$ is simply $\overline{CAP}_O(S)$ with $S = O$. Since in that case $s_{ij} = o_{ij}$, it is clear that if o_{ij} is big, s_{ij} is too. Hence, of all the possible values for $\overline{CAP}_O$ for synthetic datasets, $\overline{CAP}_O(O)$ will be one of the larger ones. In other words, almost any randomly selected synthetic dataset will have an average CAP smaller than the original dataset.

One may also note that as O gets more *informative*, meaning that there is a clearer relationship between the keys and the targets, some cells of the original dataset will get bigger and the others ones will get smaller, so that as O gets more informative his average CAP value will increase. As a result, any dataset such as O_2, where keys perfectly predict the target, will reach the maximum value of $\overline{CAP}_O(\text{O}) = 1$.

5 Discussion

Our careful study of the CAP risk measure proposed in [1] generated several interesting results. We first described how by construction improving the value of the CAP measure for some individuals usually worsened it for others in the datasets. We also showed that to minimize the value of the average CAP measure one could use a synthetic dataset with a single non-empty cell, the choice of which depends on the version of the measure used. We then observed empirically a few limits of the CAP average measure, namely that a low average CAP may hide some observations with large disclosure risk, that the attribution probabilities used in CAP do not necessarily reflect the risk posed by intruders using the synthetic dataset to predict targets for individuals, and that most synthetic datasets will have an average CAP value smaller than the original dataset.

There remains some aspects of the measure which we did not investigate in the paper. For example, [1] also propose evaluating the risk associated with a synthetic dataset S by comparing $\overline{CAP}_O(S)$ to a certain baseline CAP, which is relative to the knowledge that an intruder would have about the target regardless of the released dataset, in particular, the univariate distribution of the variable in the population. In future work, we would like to explore how this comparison measure behaves with different type of datasets.

Future work should also study other proposed measures of risk for synthetic datasets to verify to which extent they share some of the limits of the average CAP risk measure. Understanding what kind of synthetic datasets they favor, for example, would help understand better the difference between these measures of risk.

Finally, this works suggest that we should reflect some more on the need and intent behind protecting inferential disclosure from synthetic datasets. While one may reasonably argue that re-identification risk should be minimized, if only to ensure trust from the respondents, the situation is not so clear for inferential disclosure. The problem is that there is a very thin line between inference and inferential disclosure. For example, consider the original dataset O_2 again. Because of the perfect association between keys and targets, individuals have a very high inferential disclosure risk. But even releasing a model fitted on the dataset might reveal as much information about the individuals. Hence, one might want to reconsider whether this constitutes inferential disclosure, or whether this is simply the intended inference from collecting the data. In fact, even differential privacy [11], which is often seen as a very conservative measure of risk, may allow this type of inferential disclosure if the original data is very informative about the variables of interest. In other words, we believe that the intent of the original dataset, and the types of inferences expected by the respondent, should inform our evaluation of disclosure risks for the respondents, and not simply technical measures of inferential disclosure risk. We consider this an important avenue for future research in statistical disclosure control.

Bibliography

[1] Taub, J., Elliot, M., Pampaka, M., Smith, D.: Differential correct attribution probability for synthetic data: An exploration. In: International Conference on Privacy in Statistical Databases, Springer (2018) 122–137

[2] Abay, N.C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., Sweeney, L.: Privacy preserving synthetic data release using deep learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer (2018) 510–526

[3] Jordon, J., Yoon, J., van der Schaar, M.: Pate-gan: Generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations. (2018)

[4] Liu, F.: Model-based differentially private data synthesis. arXiv preprint arXiv:1606.08052 (2016)

[5] Frikken, K.B., Zhang, Y.: Yet another privacy metric for publishing micro-data. In: Proceedings of the 7th ACM workshop on Privacy in the electronic society. (2008) 117–122

[6] Hu, J.: Bayesian estimation of attribute and identification disclosure risks in synthetic data. arXiv preprint arXiv:1804.02784 (2018)

[7] Nin, J., Herranz, J., Torra, V.: Using classification methods to evaluate attribute disclosure risk. In: International Conference on Modeling Decisions for Artificial Intelligence, Springer (2010) 277–286

[8] Elliot, M., Domingo Ferrer, J.: The future of statistical disclosure control. Paper published as part of The National Statistician's Quality Review (2018)

[9] Elliot, M.: Final report on the disclosure risk associated with the synthetic data produced by the sylls team. Report 2015 **2** (2015)

[10] Hittmeir, M., Mayer, R., Ekelhart, A.: A baseline for attribute disclosure risk in synthetic data. In: Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy. (2020) 133–143

[11] Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference, Springer (2006) 265–284