Raphaël Peschi, François-Xavier Standaert*, and Vincent Blondel

From Minimal Distortion to Good Characterization: Perceptual Utility in Privacy-Preserving Data Publishing

Abstract: Definitions of utility in databases are usually torn between two main options. On the one hand, specific metrics quantify utility based on the purpose of the collected data, which hardly allows comparing the utility of data collected for different purposes. On the other hand, general purpose metrics rely on the goal of minimal distortion, which only guarantees that any privacy-preserving operation has only degraded the original data to a limited extent, and therefore does not relate to the actual utility of the original data. In this paper, we introduce an alternative solution to measure "perceptual utility", based on whether the collected data represents well the true distribution of the random variables from which it is sampled. Intuitively, perceptually useful data can be seen as data that is "useful for anything". It can therefore be connected to emerging discussions about regulations on the automatic processing of (personal) data. For this purpose, we first define this notion and show that it allows shedding interesting light on the tradeoff between utility and *anonymity* in databases. We then put forward that perceptual utility can be seen as another type of *privacy* metric, and discuss its conceptual differences and links with anonymity metrics.

1 Introduction

The collection of digital information by organizations has been an increasingly important trend over the last decade. While it creates new opportunities for knowledge-based decision making, it also raises new challenges regarding the privacy of the individuals whose personal information is collected. In this context, privacy-preserving data publishing aims at releasing data in a way that it is practically useful (e.g. allows data mining) while preserving individual privacy. In other words, it aims at trading utility and privacy.

A wide literature has investigated metrics for privacy, including the popular k-anonymity [12] and its numerous refinements. By contrast, metrics for measuring the utility of a database are sparser, and generally face the difficulty of defining what is useful data. In order to be independent of the type of data processing purposed, the typical solution is to follow the "principle of minimal distortion". This means assuming that the database is useful anyway, and measuring a utility based on this a priori, by quantifying the damage caused by the anonymization of the data [7]. Quite naturally, this approach also implies that any modification of the original data is damaging by definition/assumption.

In this paper, we aim to investigate an alternative track for measuring utility, based on recent advances in the certification of the information leakage in cryptographic implementations [4, 6]. More precisely, we propose to quantify utility based on whether the statistical attributes of which the samples form a database are "well characterized". We further describe how to use the notion of Perceived Information (PI) for this purpose. Intuitively, the PI captures the amount of information that an adversary can extract from some observations, given a (possibly biased) model of the data. If the model is perfect, the PI correspond to Shannon's classical definition of Mutual Information (MI). If the model is imperfect (as usually the case in practice), the PI is the best approximation of the MI that is available to the adversary. Based on this notion, we can trade the speed of convergence of a model with its informativeness, and derive a perceptual utility metric for actual databases. We use this metric to illustrate concrete situations where the anonymization of the data does not have any utility cost. For example, the accuracy of an attribute's observations can be too high for being

Raphaël Peschi: Nokia, E-mail: raphaelpeschi@gmail.com
***Corresponding Author: François-Xavier Standaert:** EPL/ICTEAM, Université catholique de Louvain, Louvain-la-Neuve, Belgium, E-mail: fstandae@uclouvain.be
Vincent Blondel: EPL/ICTEAM, Université catholique de Louvain, E-mail: vincent.blondel@uclouvain.be

characterized with the number of available samples. In this case, reducing the accuracy of the collected data is beneficial to anonymity, but does not reduce perceptual utility. We also describe a couple of experiments that allow us to discuss the impact of grouping users from the anonymity and perceptual utility points-of-view, as well as the curse of dimensionality for the characterization of attributes. Eventually, we conclude the paper by arguing why perceptual utility should also be seen as another facet of privacy, and can be asymptotically connected to some (probabilistic) anonymity metrics.

Besides, and more fundamentally, perceptual utility has strong connections with recent EU regulations on the automatic processing of (personal) data, e.g. [15], and important debates regarding the impact of data collection and processing on (algorithmic) governmentality [11]. By algorithmic govenmentality, we refer to *an unprecedent mode of government fuelled mostly with infra-personal, meaningless but quantifiable signals, addressing individuals through their "profiles" – behavioral patterns produced on a purely inductive base – rather than through their understanding and will* [10]. In this context, perceptual utility can be viewed as a natural (general purpose) metric allowing a quantitative discussion of the risks of discrimination in big data systems.

Notations. In the following, we use capital letters for random variables, small caps for their samples, calligraphic letters for sets and sans serif fonts for functions.

2 Definitions and framework

In this first section, we introduce the definitions and mathematical framework that allow us to reason formally about privacy and utility in databases.

We start by defining a set of n *users* $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$, and m random variables $X_1, X_2, \ldots, X_m$ with (discrete or continuous) sample spaces $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_m$. We denote these random variables as *attributes*, that are specified by their probability function $\mathsf{p}(x)$ in the discrete case, and probability density function $\mathsf{f}(x)$ in the continuous case. We also use the more generic term probability distribution to denote both $\mathsf{p}(x)$ and $\mathsf{f}(x)$, when we do not want to distinguish between discrete and continuous random variables.

Deterministic data. We define a deterministic *Data Structure* (DS) as a set of m attributes together with their probability distributions $\mathsf{p}_i(x)$ or $\mathsf{f}_i(x)$, with $1 \leq i \leq m$. And we define a deterministic *Data Base* (DB) as the sampling of a deterministic DS, i.e. a set of $m \times n$ samples, one per attribute X_i and user u_j.

An illustration of deterministic DB is given in Table 1, where $x_i(u_j)$ denotes the sample corresponding to attribute i for user j. Concretely, purely deterministic attributes are not so frequent (the date of birth is a typical example). By contrast, there are many attributes that are stable enough for being considered as deterministic (such as the ZIP code for example), and therefore will be interpreted as such in practice.

Table 1. Example of deterministic DB.

	X_1	X_2	...	X_m
u_1	$x_1(u_1)$	$x_2(u_1)$	...	$x_m(u_1)$
u_2	$x_1(u_2)$	$x_2(u_2)$	...	$x_m(u_2)$
...	...	...	...	...
u_n	$x_1(u_n)$	$x_2(u_n)$	...	$x_m(u_n)$

Probabilistic data. Similarly, we define a probabilistic DS as a set of m attributes, each of them described by n probability distributions $\mathsf{p}_i^j(x)$ or $\mathsf{f}_i^j(x)$, with $1 \leq j \leq n$. And we define a probabilistic DB as the sampling of a probabilistic DS, where we denote the number of samples per user as $m \times n_{u_j}$, so that the total number of samples in the DB equals $m \times \sum_{j=1}^{n} n_{u_j}$.

An example of probabilistic DB with $n_{u_j} = 2$ for all users is given in Table 2, where $x_i(u_j, t)$ denotes the t^{th} sample corresponding to attribute i for user j. Concretely, any (shopping, cultural, ...) user preference can be examples of probabilistic attributes. The latter ones are especially important for our following discussions, since perceptual utility typically aims at quantifying the accuracy and convergence of their characterization.

Table 2. Example of probabilistic DB.

	X_1	X_2	...	X_m
u_1	$x_1(u_1, 1)$	$x_2(u_1, 1)$	...	$x_m(u_1, 1)$
	$x_1(u_1, 2)$	$x_2(u_1, 2)$	...	$x_m(u_1, 2)$
u_2	$x_1(u_2, 1)$	$x_2(u_2, 1)$	...	$x_m(u_2, 1)$
	$x_1(u_2, 2)$	$x_2(u_2, 2)$	...	$x_m(u_2, 2)$
...	...	...	...	...
u_n	$x_1(u_n, 1)$	$x_2(u_n, 1)$	...	$x_m(u_n, 1)$
	$x_1(u_n, 2)$	$x_2(u_n, 2)$	...	$x_m(u_n, 2)$

Both for deterministic and probabilistic data, we further denote any vector of m samples corresponding to a line of a DB (excluding the user) as an *observation.*

We also define any subset of users that one may be interested to characterize as a *group*, and a set of q groups as $\mathcal{G} = \{g_1, g_2, \ldots, g_q\}$. Eventually, we call aggregation the process of replacing an attribute X_i by an *aggregated attribute* Y_i, such that the original sample space $\mathcal{X}_i$ is replaced by a set of events $\mathcal{Y}_i$, with $|\mathcal{Y}_i| < |\mathcal{X}_i|$ if the attribute was discrete, and $\mathcal{Y}_i$ a discretized version of $\mathcal{X}_i$ if the attribute was continuous. Note that in concrete case studies the DS is always unknown and the only thing that can be analyzed are sampled DB.

3 Privacy metric(s)

As mentioned in introduction, numerous metrics were introduced to quantify various aspects of privacy in databases, some of them being surveyed in [7]. For simplicity, we will mostly use the popular k-anonymity. In this respect, and while other metrics offer stronger guarantees, we recall that our goal is not to argue about the relevance of one or another anonymity metric. We just use the k-anonymity to put forward intuitive differences between the general concepts of anonymity, privacy and utility (independent of the chosen metric). For this purpose, we first denote an observation o_j^t as the vector of t^{th} samples obtained for the m attributes of user u_j as:

$$o_j^t := [x_1(u_j, t), x_2(u_j, t), \ldots, x_m(u_j, t)]\cdot$$

Secondly, we denote the set of observations $\mathcal{O}(u_j)$ found in a DB for a user u_j as:

$$\mathcal{O}(u_j) := \{o_j^t \mid 1 \leq t \leq n_{u_j}\}\cdot$$

Thirdly, we denote the anonymity set $\mathcal{A}(o)$ as the set of users for which a given observation o is in the DB as:

$$\mathcal{A}(o) := \{u_j \mid o \in \mathcal{O}(u_j)\}\cdot$$

Based on these notations, we say that a DB preserves k-anonymity (or is k-anonymous) if:

$$k = \underset{o \in \text{DB}}{min} |\mathcal{A}(o)|\cdot$$

Intuitively, the k-anonymity guarantees that an observation does not allow to (strictly) distinguish (i.e. with probability one) a user from at least $k-1$ other users in the DB. Concretely, this metric is usually computed with respect to deterministic attributes that are supposed to be easier to collect for the adversary (such as the sex, ZIP code, ...) in order to obtain sensitive information (such as the incomes, medical data, ...).

Note that in probabilistic DB, the k-anonymity ignores the possibility that different users have different probabilities given an observation. Yet, by denoting the number of apparitions of an observation o in a DB as $\#o$ and its number of apparitions for user u_j as $\#o|u_j$, we can additionally define the *hypothetical probability* of a user u_j given an observation o as follows:

$$\tilde{\Pr}[U = u_j | O = o] := \frac{\#o|u_j}{\#o}\cdot$$

The term hypothetical probability here reflects the fact that $\tilde{\Pr}[u_j|o]$ is defined based the sampled data of a DB, which does not mandatorily represents well the DS (i.e. the true distribution of the attributes). Such hypothetical probabilities can then be used to estimate other anonymity metrics such as the one of Diaz et al. in [2], that we will use in Section 7 to illustrate the fact that perceptual utility reflects another facet of privacy.

4 Perceptual utility metric

Approaches to guarantee privacy in DB generally imply a number of anonymization operations, which include aggregation, noise addition, suppression, ... This leads to the problem of determining if the sanitized data remains useful. Both general purpose and specific metrics have been introduced to answer this question [1].

On the one hand, general purpose metrics rely on the goal of minimal distortion. That is, they start from the a priori that the DB is useful, and quantify (pseudo) utility by measuring the distance between the original and anonymized DB. To a good extent, this approach resembles the one to quantify privacy in the previous section, since it is also based on the sampled data of a DB, independent of its DS. We use the term *pseudo utility* to reflect this fact. Minimizing distortion does not guarantee that an anonymized DB is useful, it only guarantees that it is nearly as useful as originally.

On the other hand, specific metrics aim at measuring utility based on the purpose of the data collected (e.g. estimating some statistical moment for an attribute, or classifying users based on some machine learning tool). Compared to the previous case, this approach suffers from the complementary drawback that it does not allow comparing the utility of data collected for different purposes. Moreover, it requires knowing this purpose precisely at the time the data is published.

Strictly speaking, this last drawback seems unavoidable: utility is indeed most accurately defined in function of a task to perform. However, we argue next that an alternative path is possible, that may better suit current trends in big data systems. Namely, we propose to quantify (perceptual) utility based on whether the data collected represents well the DS (i.e. the true distribution of the attributes). We first define the perceived information metric that we will use for this purpose, and then provide the rationale behind our new approach.

4.1 The Perceived Information

The Perceived Information (PI) was introduced in the context of side-channel attacks against cryptographic devices, of which the goal is to recover some secret data (aka key) given some physical leakage [9, 14]. The PI aims at quantifying the amount of information about the secret key, independent of the adversary who will exploit this information. Informally, we will use this metric in a similar way, by just considering the users' ID as the secret to recover, and the observations as leakages.

Using the previous notations, we can first define the Mutual Information (MI) between the users random variable U and the observation random variable O:

$$\mathrm{MI}(U;O) = \mathrm{H}[U] + \sum_{u} \Pr[u] \cdot \sum_{o} \mathsf{p}(o|u) \cdot \log_2 \Pr[u|o],$$

if the observations are discrete, and:

$$\mathrm{MI}(U;O) = \mathrm{H}[U] + \sum_{u} \Pr[u] \cdot \int \mathsf{f}(o|u) \cdot \log_2 \Pr[u|o] \; do,$$

if they are continuous. For conciseness, we use the notation $\Pr[X = x] := \Pr[x]$ when clear from the context. The probability $\Pr[u|o]$ is derived via Bayes' theorem, e.g. $\Pr[u|o] = \frac{\mathsf{f}(o|u)}{\sum_{u^*} \mathsf{f}(o|u^*)}$ for the continuous case, and $\mathrm{H}[U]$ is computed based on the a priori distribution of the users (e.g. $\mathrm{H}[U] = \log_2(n)$ if it is uniform).

Concretely, and as previously discussed, the true distribution of the attributes (i.e. the DS) is always unknown. Therefore, it is not possible to compute the MI directly (excepted in the case of simulated DB). In order to avoid this caveat, the approach in side-channel analysis, that we repeat here, is to split the DB that one wishes to evaluate in two parts: the first one, denoted as $\mathrm{DB_l}$ is used for learning a model, the second one, denoted as $\mathrm{DB_t}$ is used to test its accuracy.[1]

The PI is then computed based on two main phases:

1. A probabilistic model $\hat{\mathsf{p}}^j_{\mathrm{model}}$ (resp. $\hat{\mathsf{f}}^j_{\mathrm{model}}$) is estimated for each user u_j, which we denote with the conditional distribution $\hat{\mathsf{p}}^j_{\mathrm{model}}(o|u_j) \leftarrow \mathrm{DB_l}$ in the discrete case (resp. $\hat{\mathsf{f}}^j_{\mathrm{model}}(o|u_j) \leftarrow \mathrm{DB_l}$ in the continuous case).

Note that in the discrete case, such a model can be quite close to the previously defined hypothetical probabilities. The main conceptual difference is that this model is only built from a (learning) part of the DB that will be tested on independent observations (in the second phase below), and can be "simplified" (see, e.g. the example in Section 6.5 taking advantage of an independence assumption). By contrast in the continuous case, differences are generally more explicit, since the model will be based on a continuous distribution.

2. The model is tested by computing the PI estimate:

$$\hat{\mathrm{PI}}(U;O) = \mathrm{H}[U] + \sum_{j=1}^{n} \Pr[u_j] \cdot \sum_{k=1}^{n^t_{u_j}} \frac{1}{n^t_{u_j}} \cdot \log_2 \hat{\Pr}_{\mathsf{model}}[u_j|o_j^k],$$

where $n^t_{u_j}$ is the number of observations for user u_j in $\mathrm{DB_t}$, and $\hat{\Pr}_{\mathsf{model}}[u_j|o_j^k]$ is derived from $\hat{\mathsf{p}}^j_{\mathrm{model}}$ (resp. $\hat{\mathsf{f}}^j_{\mathrm{model}}$) via Bayes' theorem, as for the standard MI.

In the ideal case where the model is perfect, the PI is an estimate of the MI (i.e. its value tends towards the MI one as the number of samples in $\mathrm{DB_t}$ increases). In the practical cases where the model differs from the attributes' true distribution, the PI captures the amount of information that is extracted from the DB, biased by the model errors. That is, the PI becomes lower than the MI as the model errors increase. It can even become negative in contexts where the model does not approximate at all the attributes' true distribution.

Note that the PI can be viewed as a general purpose utility metric in both cases, since it aims at characterizing a distribution independent of the goal of the data collection. Yet, the interpretation of this general purpose flavor is only straightforward in the ideal case (where the model perfectly corresponds to the DS). By contrast, in the practical cases where the PI differs from the MI, it then requires to analyze this difference (i.e. is it due to estimation or assumption errors?). Interestingly, we argue next that one can directly take advantage of leakage certification tools for this purpose.

4.2 Perceptual utility rationale

In the following, we will say that a DB is more or less *perceptually useful* if it allows to extract a smaller or larger (positive) amount of PI on its users. As previously

[1] One can possibly split the DB in more parts in order to take advantage of k-fold cross-validation, as described in [6].

mentioned, the word perceptual relates to the fact that the definition of the PI is based on a model for the DS attributes, which may be incorrect (because of estimation or assumption errors). Intuitively, a perceptually useful DB implies that the collected data represents the DS sufficiently well to capture some specific features of the users which allow to distinguish them (i.e. are useful for anything). Hence it is conceptually different from metrics based on the hypothetical probabilities of a DB, and will be concretely different whenever the number of samples in a DB is too low for characterizing the attributes accurately or if the model is based on some (possibly incorrect) assumptions on the underlying distribution. Furthermore, we will say that a DB is *perceptually characterized* with respect to some modeling tool if the PI metric computed based on this tool has converged (i.e. is stable over the number of observations in the DB). Eventually, we will say that a DB is *perfectly characterized* with respect to some modeling tool if it is perceptually characterized and the model used when estimating the PI does not exhibit detectable assumption error (with leakage certification tests – see next).

Note that besides the fact that it captures possible estimation and assumption errors, the word perceptual additionally relates to the difficulty to characterize users in the long term, e.g. because of preference or habit changes. In other following, we will typically assume that the users' distributions are stationary.

The rationale behind the PI metric and the previous notions of utility relate to two recent results in the field of side-channel attacks against cryptographic devices:

1. The perceived information can be used to bound the success rate of a Bayesian adversary trying to distinguish a user based on new samples of his DS (i.e. independent of the samples used to build the model) [4]. This is in contrast with the k-anonymity game, where the goal is to identify a user based on an observation that is already in a DB. In other words, it relates to the "best possible" characterization of the attributes that can be obtained thanks to statistical sampling, and therefore to the possibility to effectively discriminate the DB users based on this characterization. This justifies why the PI is an interesting candidate metric for quantified discussions regarding algorithmic governmentality.

2. The perceived information can benefit from "leakage certification" [6], which aims to guarantee that it is "close enough" to the MI. Informally, we say that the PI is close enough to the MI if any discrepancy between these metrics is dominated by assumption errors. More precisely, leakage certification guarantees that, for a given number of observations in a DB, any improvement of the modeling tool (used to characterize an attribute and compute $\hat{\Pr}_{\mathsf{model}}[u_j|o_j^k]$) will not lead to a significant increase of the PI since the assumption errors are anyway smaller than the estimation ones for this number of observations. Leakage certification tools are based on two steps. First an analysis of estimation errors which allows determining whether a DB is perceptually characterized. Second an analysis of assumption errors which allows determining whether a DB is perfectly characterized. If this last condition is satisfied, the collected data is "useful for everything" since it nearly perfectly represents the true attributes' distribution.

We refer to these previous works for the bounds' proof and details on the implementation of certification tools. Note that our notion of perceptual utility is based on whether some users' distributions are well characterized. This directly corresponds to the setting of a probabilistic DB (where each attribute is indeed distributed according to some unknown distribution). However, even in the case of deterministic DB, one will generally describe group features, in which case the deterministic user data also becomes probabilistic (e.g. the year of birth is a deterministic attribute for single users, but becomes a probabilistic attribute for groups). In this respect, by useful data, we essentially mean data that is useful to discriminate a group or an individual.

4.3 Related works

The use of information theoretic metrics to quantify the tradeoff between utility and privacy in statistical DB is admittedly not new: see e.g. [3]. In general, statistical distance metrics (such as the MI) are indeed natural candidates to measure a generic dependency between different random variables. However, our approach fundamentally differs from such a previous work since (as already observed in Section 3 for privacy metrics) it aims at quantifying model errors rather than assuming a model is perfect and considering hypothetical probabilities. That is, our work is primarily concerned with the question whether a DB reveals something about its users' true distributions and therefore whether it can be discriminating (e.g. for decision making) and how much. By contrast, [2] and [3] consider the DB independent of the number and representativity of the samples it contains. For a similar reason, comparisons between general purpose utility metrics based on quantifying the distortion between a DB and its sanitized version and perceptual utility are essentially meaningless since they quantify different things: how much sanitization has modified a DB in the first case; how much it has decreased the representativity of its samples in the second one.

4.4 Additional remarks

We mention that the perceptual utility in this section is specified based on the conditional probability $\hat{\Pr}_{\mathsf{model}}[u|o]$ and as a reduction of the user's entropy, which typically captures the risk of re-identification of a user based on his observations. But perceptually utility could be similarly defined based on the possibility to predict the observations (i.e. preferences) of a given user. We also note that the PI is an average metric, which is convenient to analyze a DB globally. However, formally the link between the PI metric and the success rate of a Bayesian adversary only holds per user, which can be analyzed by computing the PI per user $\hat{\mathrm{PI}}(u;O)$.

5 Preliminaries

We now present two simple observations that result from the previous definitions and framework, and will support our simulated experiments in the next section.

First, the **impossibility of privacy preserving data publishing for continuous attributes** is formalized by the next theorem (a proof sketch is in Appendix A).

Theorem 1. *In a (finite) DB sampled from a deterministic or probabilistic DS with a single continuous attribute, every observation is different with probability 1.*

It implies that every user can be uniquely identified based on a single observation in such a DB and therefore that no k-anonymity can be guaranteed unless some preliminary sanitization of the data is applied (which necessarily implies some kind of discretization step).

Next, the **unicity of discrete attributes from the privacy metrics point-of-view** is formalized next:

Theorem 2. *Any DS with discrete attributes $X_1, ..., X_m$ can be represented as a DS with single discrete attribute X and having the same k-anonymity.*

The result simply relies on defining an attribute X of which the sample space is the product of the sample spaces of the attributes $X_1, ..., X_m$. Note that a similar statement holds for the characterization point-of-view, as long as one tries to characterize all the attributes jointly. However, as will be discussed next, it is sometimes useful to consider them independently to reduce the sampling complexity of the characterization.

6 Simulated experiments

In this section, we analyse the evolution of the k-anonymity and perceptual utility in the exemplary case of a simulated database containing individuals' shopping lists. We first define our simulation settings. Next, we put forward a number of intuitions regarding the impact of aggregating attributes, grouping users and the list's curse of dimensionality on our metrics.

Note that despite considering simulated experiments, the only thing we use to compute our metrics are sampled DB as it would be the case in concrete case studies. The main difference between simulated and concrete experiments is that we can generate DB of various sizes, which allows us to generate enough samples for our metrics to converge (which helps readability). Also, since we actually know (i.e. choose) the exact distribution of the DS, we know whether a DB is perfectly characterized without the need to run the second (assumption error) part of leakage certification tests.[2]

6.1 Simulation settings

We consider a DS (corresponding to a shop) with n users (aka clients). The shop sells N_i different items. For simplicity, each item can be puchased in N_q (integer) quantities. Hence, we have a set of $N_l = N_q^{N_i}$ possible shopping lists defined as:

$$\mathcal{L} = \{(q_{i_1}, q_{i_2}, \ldots, q_{i_{N_i}})|1 \leq q_{i_j} \leq N_q\},$$

that we will also denote as $\mathcal{L} = \{l_1, l_2, \ldots, l_{N_l}\}$. For example, a shop with $N_i = 2$ items and $N_q = 3$ quantities will lead to the following set of $3^2 = 9$ lists:

$$\{(0,0),(0,1),(0,2),(1,0),(1,1),(1,2),(2,0),(2,1),(2,2)\}.$$

In this context, each user u_j has a single attribute. We define our simulated DS by selecting the user's probability functions $\mathsf{p}_1^j(o|u_j)$. For this purpose, for each user we

[2] Which is the most expensive part of the certification process, although recent works have made steps to simplify it [5].

pick up N_l probabilities randomly. (Technically, we did that by assigning a random Gaussian-distributed sample to each list, and by normalizing to so that the sum of the list probabilities equals one for every user. Adapting the variance of the Gaussian distribution additionally allowed us to make users more or less different). Concretely, we analyzed a case study with $n = 100$ users, $N_i = 4$ items and $N_q = 5$ quantities (i.e. $5^4 = 625$ possible lists). The number of observations per user varies in our experiments, but it is always identical for all users. Eventually, and taking advantage of our simulated context, we report results averaged over 100 sampled DB (which allows obtaining smoother curves and gaining intuition about the average behavior of our metrics).

6.2 PI metric estimation

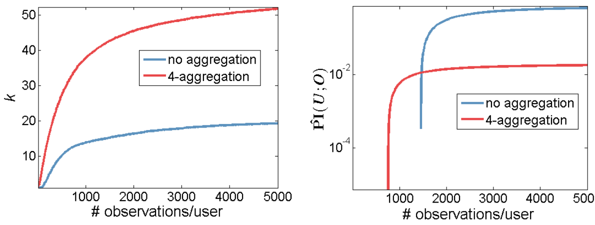
All our metrics are estimated in function of the DB size, quantified with a number of observations per user. For a given number of observations per user, we compute the PI as described in Section 4.1, taking advantage of 10-fold cross validation as suggested in Footnote 2. That is, we iteratively take $\frac{9}{10}$ of the observations to build a model for all the users, and use the last $\frac{1}{10}$ of the observations to test it (i.e. produce the estimated probabilities $\hat{\Pr}_{\mathsf{model}}[u_j|o_j^k]$) and compute the PI. Since our simulated case study is based on a discrete attribute, the models simply correspond to histograms capturing the probabilities of occurence of the lists. Note that in practice, the size of a DB is generally fixed and one can typically not ask for more observations. Yet, even in this case the evaluation of perceptual utility benefits from estimating the PI with gradually increasing subsets of the DB in order to gauge the models' convergence. Finally, and for readability, all our results are provided as plots of the metrics. Since they are based on a simulated case study, the actual values observed in the experiments are not relevant and only used to discuss the intuitions behind our new approach to (perceptual) utility.

6.3 Impact of aggregation

As a first illustration of the tradeoff between k-anonymity and perceptual utility, we investigate the impact of a simple aggregation process for the previously defined shopping list attribute. Namely, we define N_a-aggregated shopping lists as lists where sets of N_a original (consecutive) items are considered as single (aggregated) items that can be purchased in $N_q' = N_a \cdot (N_q - 1) + 1$ quantities. This reduces the cardinality of the set of lists from $N_q^{N_i}$ down to $N_q'^{N_i/N_a}$. For simplicity, we only consider cases wher N_a divides N_i.

Figure 1 represents the evolution of the k-anonymity and the PI in function of the size of the DB, for the 100-users DB defined in Section 6.1, with and without aggregation. By making users more similar, the aggregation process increases the k-anonymity and *asymptotically* decreases the PI. But quite interestingly, we see that for a DB with up to 1500 observations per user, the PI of the aggregated data is in fact larger than for the orginal one. This typically corresponds to the "win-win" scenario mentioned in introduction. That is, the amount of data collected is not sufficient to fully characterize the original lists. So aggregation allows improved k-anonymity without any loss of (perceptual) utility, because the DB without aggregation is not yet perceptually characterized and is in fact even less useful than the DB with aggregation for this number of observations.

Fig. 1. Average impact of aggregating items.



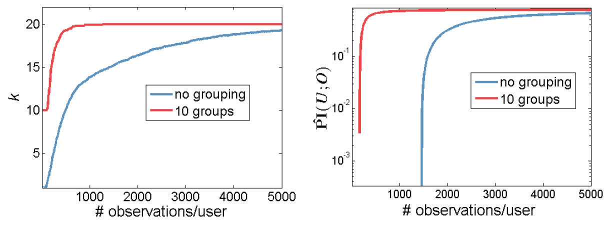
(a) k-anonymity. (b) Perceived Information.

6.4 Impact of grouping

We now study a complementary experiment in which some users are grouped together. For this purpose, and in order for the grouping to make sense, our DS described in Subsection 6.1 actually embeds an additional feature. Namely, we only created $q = 10$ user's probability functions p_1^j (with $1 \leq j \leq q$), and each of them was repeated 10 times to obtain $n = 100$ users. In this context, one can naturally group each subset of 10 identical users together. As illustrated in the right part of Figure 2, this improves the convergence of the PI metric (since we have 10 times more observations per user).
Furthermore, if grouping is perfect (i.e. users in each group have the same distribution), there is no PI loss since in our experiments with perfect models, we have:

$$\mathrm{MI}(U;O) = \log(100) + \sum_{u=1}^{100} \frac{1}{100} \sum_o \mathsf{p}(o|u) \log \Pr[u|o];$$

Fig. 2. Average impact of perfect grouping.



(a) k-anonymity. (b) Perceived Information.

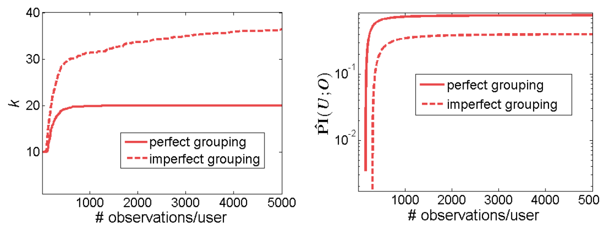
$$
\begin{aligned}
\mathrm{MI}(U;O) &= \log(100) + \textstyle\sum_{g=1}^{10} \frac{10}{100} \sum_o \mathsf{p}(o|g) \log \frac{\Pr(g|o)}{10};\\
&= \log(10) + \textstyle\sum_{g=1}^{10} \frac{1}{10} \sum_o \mathsf{p}(o|g) \log \Pr(g|o);\\
&= \mathrm{MI}(G;O).
\end{aligned}
$$

So the gap between the PI curves in Figure 2 is only due to a lack of samples to characterize ungrouped users. In other words, both characterizations become perfect once having access to a sufficient number of observations.

As for the k-anonymity in the left part of the figure, it is positively impacted by grouping as well. Indeed, whenever grouping, any observation recorded for a user u_j will be only be labeled as belonging to a group g_j. So in the simple case where groups have identical sizes that we consider, we can directly derive the user k-anonymity by multiplying the group k-anonymity by the group size. This implies a minimum k-anonymity of 10.

Quite naturally, the situation differs when the grouping is imperfect, as reflected in Figure 3. In this case, where the users in each group have different distributions, the characterization is still faster. However, it comes at the cost of a perceptual utility loss (i.e. an asymptotically imperfect characterization), which can be explained by the distributions of the groups that are becoming more similar (due to their more different users), as also reflected in a larger k-anonymity.

Fig. 3. Average impact of imperfect grouping.



(a) k-anonymity. (b) Perceived Information.

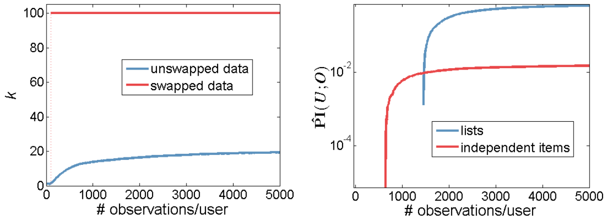
Note that the imperfect characterization due to imperfect grouping is obvious in Figure 3 since we can compare it with a perfect characterization (thanks to our simulated setting). However, even in a concrete case study where a perfect characterization would not be accessible, the fact that groups are imperfect would be directly detected thanks to leakage certification.

Let us additionally mention that grouping is a relevant option to preserve (generalizations of) k-anonymity when multiple observations for probabilistic attributes are leaked for a single user (since their combination usually allows a much better discrimination), which we leave as an interesting scope for further research.

6.5 The curse of dimensionality

As clear from the previous discussions, the size of the sample space for shopping lists' distributions grows exponentially in the number items they contain. This suggests that exhaustively characterizing such lists rapidly turns out to be infeasible (despite our toy examples made it possible by limiting N_i to 4). In this context, a last natural direction, that we investigate in this subsection, is to characterize items independently. As illustrated in the right part of Figure 4, this allows making the collected data perceptually useful much faster. We further observe that the independence assumption was incorrect in our setting, since asymptotically the characterization of four independent items is significanly less informative than the characterization of full lists. Again, this type of incorrect assumption would be directly detected by applying leakage certification tools.

Fig. 4. Average impact of independent items characterization.



(a) k-anonymity. (b) Perceived Information.

This last example suggests that in practical case studies, the PI may significantly differ from the MI because of the difficulty to characterize large distributions in a non-parametric manner. It typically happens in contexts where a DB can only be analyzed based on some simplifying assumptions. Note that even in this case, perceptual utility remains a general purpose metric in the sense that making assumptions about a distribution to maximize the PI is different than deciding in advance the goal for which some data is collected. Importantly, this observation also highlights that the risks of discrimination in big data systems are inherently hard to bound when exploiting the collected data requires doing (non-quantitative) a priori assumptions on the structure of this data (that are hard to analyze systematically).

Interestingly, and assuming that only single items have to be characterized, it also becomes possible to sanitize a DB with "utility-preserving" operations that substantially increase the k-anonymity. In particular, it is easy to see that some types of data swapping (similar to the proposal in [8]) will not affect the utility of items considered independently. For example, let us assume 3 user observations o_1, o_2, o_3 made for 3 items i_1, i_2, i_3. In this case, any permutation of the lines below will lead to "potential user observations" o'_1, o'_2, o'_3 that do not modify the independent items' characterization.
This example is illustrated with the sampled DB below, where the swapping does not modify the distribution of the items (while it does modify the list distribution):

	o_1	o_2	o_3			o'_1	o'_2	o'_3
i_1	1	0	1	swap →	i_1	0	1	1
i_2	0	0	1		i_2	0	1	0
i_3	2	1	1		i_3	2	1	1

Since the permutations are unknown, such operations increase the number of potential user observations, and therefore the k-anonymity. Quite naturally, this also makes the computation of the k-anonymity more challenging, but it at least guarantees that as soon as every quantity appeared once for every user and item, the k-anonymity will be maximum. As illustrated in the left part of Figure 4, this condition was typically observed after 50 observations per user in our case study.[3] As the previous grouping, this type of anonymization will preserve k-anonymity even in contexts where multiple observations are leaked about a user. But contrary to grouping, it will not maintain probabilistic anonymity metrics such as the anonymity degree in [2].

7 Utility is privacy (loss)

In the present state-of-the-art, privacy and utility are usually seen as two different and conflicting goals. However, the results in this paper suggest that this intuition highly depends on the size of the DB, leading to two important observations. First, there are examples where privacy metrics such as the k-anonymity can be improved without loss (or even with an improvement) of perceptual utility. They typically corresponds to situations where the size of the DB is too small to characterize the attributes' true distribution. Second and maybe more fundamentally, the most striking conclusion of our experimental case studies is that, as the number of samples in a DB increases, both the k-anonymity and the characterization of its users generally increases. In this respect, anonymity and perceptual utility should in fact be seen as two different facets of privacy. On the one hand, anonymity allows a user to deny allegations (i.e. claiming that he is not the only one exhibiting some attributes). On the other hand, useful data characterizes its users, which potentially allows identifying them based on observations that are not (yet) in the DB. This last observation naturally suggests to investigate connections between anonymity metrics and perceptual utility, which we will do next with a last experiment.

For this purpose, we first come back to one limitation of k-anonymity mentioned in Section 3. Namely, this metric ignores the possibility that different users have different probabilities given an observation (e.g. we can have 100-anonymity in a case where an observation is generated by a user with probability 0.99 and by the 99 other users with probability 0.01). Since the characterization of the users with the PI is essentially based on a probabilistic reasoning, a first step to connect anonymity and percetual utility is to consider probabilistic anonymity metrics. As already mentioned, there are several published solutions for this purpose. We will rely on a metrics inspired from [2]. More precisely, we will consider the hypothetical probabilities of Section 3 and first define the (hypothetical) conditional entropy:

$$\tilde{\mathrm{H}}[U|O] = -\sum_{u} \Pr[u] \cdot \sum_{o} \tilde{\Pr}[o|u] \cdot \log_2 \tilde{\Pr}[u|o].$$

It measures the remaining anonymity of the users based on the (hypothetical) probabilities specified by the DB. Next, we define the Hypothetical Information (HI):

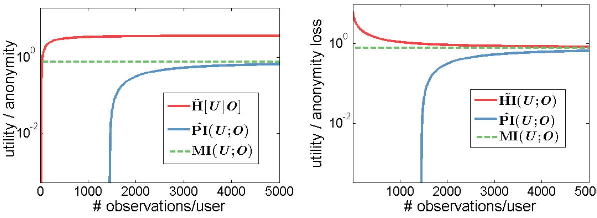
$$\tilde{\mathrm{HI}}(U;O) = \mathrm{H}[U] + \sum_{u} \Pr[u] \cdot \sum_{o} \tilde{\Pr}[o|u] \cdot \log_2 \tilde{\Pr}[u|o],$$

which similarly represents the anonymity loss of the users. Note that the anonymity degree in [2] is just the hypothetical conditional entropy normalized by $\mathrm{H}[U]$. As a result, we first computed $\hat{\mathrm{PI}}(U;O)$ and $\hat{\mathrm{H}}[U|O]$ based on the same case study as in the previous section.

[3] Note that by slightly biasing the DB with additional fake observations (which will further decrease its perceptual utility), we can enforce that this condition is met even earlier.

The result of this experiment is in the left part of Figure 5, where we additionally depicted the exact value of $\mathrm{MI}(U;O)$, which is feasible in our simulated setting. Interestingly, both the (probabilistic) anonymity and the perceptual utility again increase with the number of observations per user, which confirms that the intuitions extracted from the k-anonymity computations in the previous section can also hold for other (probabilistic) metrics such as the anonymity degree. Besides, and as expected, the PI estimate tends towards the true MI value as this number of observations per user increases, since our (discrete) models indeed tend towards the true distributions of the observations in this case.

Fig. 5. Avg. asymptotic behavior of the perceptual utility and the probabilistic anonymity (loss) based on hypothetical probabilities.



(a) PI vs. anonymity.

(b) PI vs. anonymity loss.

Next, we reported the PI together with the anonymity loss measured with $\tilde{\mathrm{HI}}(U;O)$ in the right part of the figure. This last plot reveals two essential intuitions. First, as the number of observations per user in the DB increases, the anonymity loss measured with the HI and the characterization of the users measured with the PI are getting closer. In case of a perfect modeling for discrete attributes (as in our experiments), they even tend towards exactly the same value, since the hypothetical probabilities used to compute the HI and the models used to compute the PI asymptotically tend towards the same DS distribution.[4] Second, and from a general privacy point-of-view, it illustrates that the impact of increasing the size of the DB is contrasted. On the one hand, it can improve anonymity metrics since more data can allow users to better "hide themselves" in the DB.[5] On the other hand, it also improves their characterization. In this respect, one could see probabilistic anonymity metrics (such as the HI in this paper, or Diaz et al.'s anonymity degree) as related to the *internal identifiability* of the users (i.e. the possibility to identify them based on an observation from the DB). By contrast, the perceptual utility is rather related to their *external decidability/predictability* (i.e. to whether one can use the DB to make a decision based on the user's attributes, or to predict their future observations). By collecting more data, one can reduce the internal identifiability (up to the limit given by the MI) but this comes at the cost of an improved decidability/predictability, i.e. a better characterization, and therefore with increased risks of discrimination.

8 Conclusions

In this paper, we propose to quantify the (perceptual) utility of a DB based on whether its collected data represents well the true distribution of the random variables from which it is sampled. Perceptual utility provides an alternative solution to discuss the tradeoff between anonymity and utility in DB, and captures a fundamentally different notion of utility than solutions based on the principle of minimal distortion. It provides a natural (general purpose) metric to quantify the risks of discrimination due to the automatic data processing. This metric brings a complementary view to the state-of-the-art in open data publishing since it suggests that *useful data is always worrying from the privacy viewpoint, even if it guarantees some level of anonymity.*

Our results raise a number of interesting research challenges. First, it would be relevant to study the connections between perceptual utility and the location privacy metrics in [13], for which characterization is indeed a central ingredient of the definition. Next, and for large enough DB, investigating the links between perceptual utility and the success rate of adversaries trying to identify users based on models built with a DB, that have been proved useful in cryptographic contexts, is an important open problem as well. This link could potentially improve and simplify the evaluation of location privacy. More generally, applying the tools in this paper to real case studies, and analyzing the risks when combining multiple DB, is certainly needed to confirm

[4] In case of imperfect modeling, a gap between these two asymptotic values will remain, due to the fact that the estimated PI will inevitably become lower than the (true) MI.

[5] Intuitively, this can be easily understood in the case of observations coming from distributions with large supports: indeed in this case it is likely that little observations per user create no collisions (i.e. confusion) between users, while increasing the number of observations per user can create such collisions).

their applicability and improve understanding. Eventually, developing new methods allowing one to characterize the evolution of a user's profile over time (i.e. what is the impact of a change of preferences or habits on privacy and utility?) is yet another stimulating goal.

Acknowledgments. François-Xavier Standaert is an associate researcher of the Belgian Fund for Scientific Research (FNRS-F.R.S.). This work has been funded in parts by the EU through the ERC project 724725.

References

[1] J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In Y. Li, B. Liu, and S. Sarawagi, editors, *ACM SIGKDD, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 70–78. ACM, 2008.

[2] C. Díaz, S. Seys, J. Claessens, and B. Preneel. Towards measuring anonymity. In R. Dingledine and P. F. Syverson, editors, *Privacy Enhancing Technologies, Second International Workshop, PET 2002, San Francisco, CA, USA, April 14-15, 2002, Revised Papers*, volume 2482 of *LNCS*, pages 54–68. Springer, 2002.

[3] J. Domingo-Ferrer and D. Rebollo-Monedero. Measuring risk and utility of anonymized data using information theory. In M. Mesiti, T. M. Truta, L. Xiong, S. Müller, H. Naacke, B. Novikov, G. Raschia, I. Sanz, P. Sens, D. Shaporenkov, and N. Travers, editors, *Proceedings of the 2009 EDBT/ICDT Workshops, Saint-Petersburg, Russia, March 22, 2009*, volume 360 of *ACM International Conference Proceeding Series*, pages 126–130. ACM, 2009.

[4] A. Duc, S. Faust, and F. Standaert. Making masking security proofs concrete - or how to evaluate the security of any leaking device. In E. Oswald and M. Fischlin, editors, *EUROCRYPT 2015, Sofia, Bulgaria, April 26-30, 2015, Proceedings, Part I*, volume 9056 of *LNCS*, pages 401–429. Springer, 2015.

[5] F. Durvaux, F. Standaert, and S. M. D. Pozo. Towards easy leakage certification. In B. Gierlichs and A. Y. Poschmann, editors, *CHES 2016, Santa Barbara, CA, USA, August 17-19, 2016, Proceedings*, volume 9813 of *LNCS*, pages 40–60. Springer, 2016.

[6] F. Durvaux, F. Standaert, and N. Veyrat-Charvillon. How to certify the leakage of a chip? In P. Q. Nguyen and E. Oswald, editors, *EUROCRYPT 2014, Copenhagen, Denmark, May 11-15, 2014. Proceedings*, volume 8441 of *LNCS*, pages 459–476. Springer, 2014.

[7] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4), 2010.

[8] S. P. Reiss. Practical data-swapping: The first steps. *ACM Trans. Database Syst.*, 9(1):20–37, 1984.

[9] M. Renauld, F. Standaert, N. Veyrat-Charvillon, D. Kamel, and D. Flandre. A formal study of power variability issues and side-channel attacks for nanoscale devices. In K. G. Paterson, editor, *EUROCRYPT 2011s, Tallinn, Estonia, May 15-19, 2011. Proceedings*, volume 6632 of *LNCS*, pages 109–128. Springer, 2011.

[10] A. Rouvroy. Algorihmic governmentality: a passion for the real and the exhaustion of the virtual. Transmediale – All Watched Over By Algorithms. Berlin, Germany, January 2015.

[11] A. Rouvroy. The end(s) of critique: Data-behaviorism vs. due-process. Privacy, Due Process and the Computational Turn. Philosophers of Law Meet Philosophers of Technology , M. Hildebrandt & E. De Vries (eds.), Routledge, 2012.

[12] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In A. O. Mendelzon and J. Paredaens, editors, *ACM SIGACT-SIGMOD-SIGART, June 1-3, 1998, Seattle, Washington, USA*, page 188. ACM Press, 1998.

[13] R. Shokri, G. Theodorakopoulos, J. L. Boudec, and J. Hubaux. Quantifying location privacy. In *32nd IEEE Symposium on Security and Privacy, S&P 2011, 22-25 May 2011, Berkeley, California, USA*, pages 247–262. IEEE Computer Society, 2011.

[14] F. Standaert, T. Malkin, and M. Yung. A unified framework for the analysis of side-channel key recovery attacks. In A. Joux, editor, *EUROCRYPT 2009, Cologne, Germany, April 26-30, 2009. Proceedings*, volume 5479 of *LNCS*, pages 443–461. Springer, 2009.

[15] The European Parliament and the Council of the European Union. On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC. Regulation (EU) 2016/..., April 2016.

A Proof sketch

Theorem 1. For a finite DB with a single attribute, the number of observations equals $N = \sum_{i=1}^{n} n_j$ with n the number of users and n_j the number of observations per user. In the detreministic case, (where $n_j = 1$) we must show that:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \Pr[x(u_i) = x(u_j)] = 0.$$

Since $x(u_i)$ and $x(u_j)$ are sampled from a contiuous distribution f, we have: $\Pr[x(u_i) = x(u_j)] = \int \Pr[x(u_i) = \alpha] \cdot \Pr[x(u_j) = \alpha]\ dx$, and $\Pr[x(u_i) = \alpha] = \Pr[x(u_j) = \alpha] = 0$, which concludes the proof. A similar argument holds in the probabilistic case by showing the generalized statement:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j} \Pr[x(u_i,k) = x(u_j,l)] = 0.$$