# Big Data Anonymization Requirements vs Privacy Models

Josep Domingo-Ferrer
*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, Spain*

Keywords: Privacy, Big Data, Anonymization, Privacy Models, $k$-anonymity, Differential Privacy, Randomized Response, Post-randomization, $t$-closeness, Permutation, Deniability.

Abstract: The big data explosion opens unprecedented analysis and inference possibilities that may even enable modeling the world and forecasting its evolution with great accuracy. The dark side of such a data bounty is that it complicates the preservation of individual privacy: a substantial part of big data is obtained from the digital track of our activity. We focus here on the privacy of subjects on whom big data are collected. Unless anonymization approaches are found that are suitable for big data, the following extreme positions will become more and more common: nihilists, who claim that privacy is dead in the big data world, and fundamentalists, who want privacy even at the cost of sacrificing big data analysis. In this article we identify requirements that should be satisfied by privacy models to be applicable to big data. We then examine how well the two main privacy models ($k$-anonymity and $\varepsilon$-differential privacy) satisfy those requirements. Neither model is entirely satisfactory, although $k$-anonymity seems more amenable to big data protection. Finally, we highlight connections between the previous two privacy models and other privacy models that might result in synergies between them in order to tackle big data: the principles underlying all those models are deniability and permutation. Future research attempting to adapt the current privacy models for big data and/or design new models will have to adhere to those two underlying principles. As a side result, the above inter-model connections allow gauging what is the actual protection afforded by differential privacy when $\varepsilon$ is not sufficiently small.

SCIENCE AND TECHNOLOGY PUBLICATIONS

## 1 INTRODUCTION

Big data have become a reality with the new millennium. Almost any human activity leaves a digital trace that is collected and stored by someone (sensors of the Internet of Things, social apps, machine-to-machine communication, mobile video, etc.). As a result, data from several different sources are available, and they can be merged and analyzed to generate knowledge. Big data differ from conventional data (small data) in several aspects, including their huge volume, their velocity (frequent updates or even continuous production) and their variety (they may include complex and unstructured data).

Even though big data are extremely valuable in many fields, they increasingly threaten the privacy of individuals on whom they are collected (often unawares of these). Thus, among the three dimensions of privacy (subject privacy, user privacy and owner privacy (Domingo-Ferrer, 2007)), we focus here on subject privacy and in particular in the anonymization approach to it. Statistical disclosure control (SDC, (Hundepool *et al.*, 2012)) tries to enable useful inferences on subpopulations from a data set, while preserving the privacy of the subjects to whom the records in the set correspond. Researchers have designed a good number of techniques to limit disclosure risk in data releases that refer to individual subjects, the so-called "microdata". These techniques have the common feature to keep original data secret and replace them with a modified version, which is called the *anonymized version*. In the last twenty years, on the other hand, several *privacy models* have been proposed. Instead of determining the specific transformation that must be applied to original data, a privacy model specifies a condition that, if satisfied by the anonymized data set, guarantees that the disclosure risk is kept under control. Privacy models normally have one or several parameters that determine how much disclosure risk is acceptable. Current models have been designed for a single data set, and they have several shortcomings in a big data scenario.



Domingo-Ferrer, J.
Big Data Anonymization Requirements vs Privacy Models.
DOI: 10.5220/0006830003050312
In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE 2018) - Volume 2: SECRYPT*, pages 305-312
ISBN: 978-989-758-319-3
Copyright © 2018 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

### 1.1 Contribution and Plan of this Paper

Unfortunately, SDC methods and privacy models in the literature have been designed with small data in mind, whereas the requirements of big data are different. The following extreme positions are becoming more and more common in front of the tension between subject privacy and big data analytics:

- *Nihilists*. They claim that no privacy is possible with big data. Some of them (typically governments[1]) claim that privacy needs to be sacrificed to security; others (typically corporations and some researchers[2]) claim that privacy is a hindrance to business and progress; yet others (typically Internet companies) do not make many claims but offer enticing and free services that result in privacy being overridden; finally, others claim nothing and offer nothing but gather, package and sell as many personal data as possible (data brokers (FTC, 2014)).
- *Fundamentalists*. They prioritize privacy, whatever it takes in terms of utility loss. Players in this camp are basically a fraction of the academics working in security and privacy.

While fundamentalists are unlikely to prevail over nihilists, they could make themselves more useful if they invested their research effort in finding anonymization solutions that are tailored to the natural requirements of big data.

In Section 2 of this paper, we try to identify the main requirements of big data anonymization. Then we examine how well the two main families of privacy models satisfy those requirements: we deal with $k$-anonymity in Section 3 and with differential privacy in Section 4. Seeing that neither family is completely satisfactory, we explore connections of $k$-anonymity and differential privacy with other privacy models in Section 5; we characterize those connections in terms of two principles, deniability and permutation. In Section 6 (conclusions and future research), we conclude that focusing on these two principles is a promising way to tackle the adaptation of current privacy models to big data or the design of new privacy models.

[1]In 2009 the UK Cabinet Office's former security and intelligence co-ordinator, Sir David Omand, warned: "citizens will have to sacrifice their right to privacy in the fight against terrorism". Also, in April 2016, the European Parliament backed the EU directive enabling the European security services to share information on airline passengers.

[2]E.g. Stephen Brobst, the CTO of Teradata, stated: "I want to know every click and every search that led up to that purchase... interactions are orders of magnitude larger than the transactions... the interactions give you the behavior", *The Irish Times*, Aug. 7, 2014.

## 2 BIG DATA ANONYMIZATION REQUIREMENTS

Leveraging the potential of available big data to improve human life and even to make profit is perfectly legitimate. However, this should not encroach on the subjects' privacy. Released big data should be protected, that is, transformed so that: i) they yield statistical results very close to those that would be obtained if the original big data were available, but ii) they do not allow unequivocal reconstruction of the profile of any specific subject.

SDC methods (Hundepool *et al.*, 2012) (for example, noise addition, generalization, suppression, lower and upper recoding, microaggregation and others) specify transformations whose purpose is to limit the risk of disclosure. Nevertheless, they do not prescribe any mechanism to assess the disclosure risk that remains in the transformed data. In contrast, privacy models (such as $k$-anonymity (Samarati and Sweeney, 1998) and differential privacy (Dwork, 2006), as well as $l$-diversity (Machanavajjhala *et al.*, 2007), $t$-closeness (Li *et al.*, 2007) and probabilistic $k$-anonymity (Soria-Comas and Domingo-Ferrer, 2012), among others) specify some properties to be met by a data set to limit disclosure risk, but they do not prescribe any specific SDC method to satisfy those properties. Privacy models seem more attractive, because they state the privacy level to be attained and leave it to the data protector to adopt the least utility-damaging method. The reality, however, is that most privacy models were designed to protect a single static data set and they have notorious shortcomings if used in a big data context.

For a privacy model to be useful for big data, it must be compatible with the volume, the velocity and the variety of this kind of data. To assess this compatibility, we propose to take into account to what extent the model satisfies the following properties (the last three of them described in (Soria-Comas and Domingo-Ferrer, 2015)):

- *Protection*. Anonymized big data should not allow unequivocal reconstruction of any subject's profile, let alone re-identification. While protection was also essential in anonymization of small data, it is more difficult to achieve in big data: the availability of many data sources on overlapping sets of individuals implies that a lot of attributes may be available on a certain individual, which may facilitate her re-identification.
- *Utility for exploratory analyses*. Anonymized big data that are published should yield results similar to those obtained on the original big data *for a broad range of exploratory analyses*. While utility

was also important in traditional anonymization of small data, it is more complicated to obtain useful anonymized big data, because they tend to be analyzed in ways that cannot be anticipated at the time of anonymizing them.

- *Composability*. A privacy model is composable if its privacy guarantees are preserved (possibly in a limited way) after repeated application. To put it otherwise, a privacy model is not composable if independently published data sets, each of which satisfies the model, can lead to a violation of the model when pooled together. Composability is essential for the privacy guarantees of the model to survive in a big data context, where data collection is not centralized, but distributed among several data sources. If one of the collectors cares about privacy and decides to use a specific privacy model, the guarantees of this model should be preserved (at least to some extent) after data fusion.
- *Computational cost*. This cost measures the amount of computation needed to satisfy the requirements of the privacy model. As said above, normally several alternative SDC methods are available to satisfy a certain privacy model. Thus, the cost will depend on the selected method. It is very important that the model be attainable with some efficient method, given the huge amount of big data. Ideally, the cost of the method ought to be linear or loglinear in the data set size. Methods with quadratic or superquadratic costs are not suitable. Anyway, there are options to reduce the computational cost of a method that would be too costly if directly used: one strategy is to use blocking to split the data set into smaller blocks and anonymize these separately. Obviously, blocking may have negative utility and privacy implications, so its application should be carefully pondered.
- *Linkability*. In big data, the information on an individual subject is collected from several sources. Hence, the ability to link records that correspond to the same individual or to similar individuals is basic to build big data. To preserve the privacy of subjects, each source ought to anonymize its data before releasing them. However, if each source independently anonymizes data, merging the data anonymized by different sources can become difficult if not impossible. This would reduce very substantially the variety of analyses that can be conducted on the data and, therefore, the knowledge obtainable from them. To be useful for big data, a model should allow the analyst to link independently anonymized data that correspond to the same subject. Note that, if records referred to the same subject are linked, the information on that subject is increased. This is a threat to privacy and, hence, the linkage accuracy should be lower in anonymized data sets than in original data sets.

# 3 BIG DATA ANONYMIZATION UNDER $K$-ANONYMITY

$k$-Anonymity aims to limit an attacker's ability to re-identify the record corresponding to a certain target subject. On the one hand, the data set to be protected contains *confidential attributes*, such as medical data, financial data, religion, sexual orientation, etc. (if it did not include confidential attributes, there would be no need to protect it!). On the other hand, the data set contains attributes that we will collectively name *quasi-identifier*; no value of any single quasi-identifier attribute in a record identifies its subject, but the combination of values of all quasi-identifier attributes may. For example, age, gender, zipcode and education level are a reasonable quasi-identifier, because in a certain zipcode there may be a single woman over 70 that holds a PhD. In fact, the attack model assumes that the attacker can at least identify the subject of some of the records using the quasi-identifier attributes: a way to do this is for the attacker to have access to an external database that contains the quasi-identifier attributes together with identifiers (name, passport number, etc.) for some of the subjects of the released dataset. In this case, he will manage to link the values of confidential attributes in some records of the released data set with some identifiers, which results in re-identification and undesired disclosure. To bring the re-identification probability down to $1/k$, $k$-anonymity requires that each combination of values of the quasi-identifier attributes be shared by a group of at least $k$ records in the protected data set (this group is called *k-anonymous class*).

Even in the traditional scenario of protecting a single static data set, deciding which attributes the data protector should include in the quasi-identifier is not obvious. To avoid mistakes, the protector should exactly know the attacker's background knowledge, but he does not know it.

In a big data scenario, satisfying the *protection* requirement is still more complicated, because the attacker may have a lot of background knowledge. An option to play it safe is to consider that *all attributes*, including the confidential ones, *are part of the quasi-identifier*.

Beyond the inconvenience of establishing a quasi-identifier, $k$-anonymity has another problem: al-

though it protects against re-identification, it may be insufficient to protect against attribute disclosure. This is the case when the values of a confidential attribute in the records of a $k$-anonymous class are identical or very similar. If the attacker manages to link his target subject to that class, then, even if he cannot ascertain which of the $k$ records of the class is the target subject's, he will learn the value of the subject's confidential attribute. This is also disclosure. Several $k$-anonymity extensions have been proposed to remedy this, such as $l$-diversity (Machanavajjhala *et al.*, 2007) or $t$-closeness (Li *et al.*, 2007).

In what concerns the *utility* requirement, a big data scenario in which data from many subjects are to be anonymized may allow forming $k$-anonymous classes whose subjects are more homogeneous than in a small data scenario with fewer subjects. Thus, even if exploratory analyses cannot be anticipated at the time of anonymization, *ceteris paribus* more homogeneous classes will yield better utility.

Regarding *composability*, $k$-anonymity was designed to protect a single data set and, in principle, it is not composable. If several independently anonymized $k$-anonymous data sets have been released that share some subjects, the attacker can leverage the so-called intersection attack to rule out some of the records in a $k$-anonymous class as not corresponding to the target subject. To achieve some composability, the protectors of several $k$-anonymized data sets ought to coordinate so that, for the subjects shared by two data sets, their $k$-anonymous classes contain the same $k$ subjects. In a big data environment, this coordination is not easy but it is not impossible. If it is not feasible, and/or the various data sets independently grow with time, the strategies sketched in (Domingo-Ferrer and Soria-Comas, 2016) can be used.

As to *computational cost*, $k$-anonymity is usually reached by modifying the values of the quasi-identifier attributes either by a combination of generalizations and suppressions (Samarati and Sweeney, 1998), or by microaggregation (Domingo-Ferrer and Torra, 2005). Although finding the optimal modification (to minimize information loss) results in NP problems, heuristics and blocking strategies allow reaching complexities $O(n \ln n)$ where $n$ is the number of records. Therefore, $k$-anonymity is reasonably computable for big data.

Finally, regarding *linkability*, assume we have two independently $k$-anonymized data sets that have some subjects in common. We need to see whether the records of a subject that is known to be in both files can be linked. The answer is that *at least the k-anonymous classes of the subject in both data sets can be linked.* If the two data sets share also some confidential attributes, the accuracy of the linkage can improve, because one can use the values of those attributes to link specific records within each $k$-anonymous class.

In summary, in a big data scenario, $k$-anonymity offers a composable privacy guarantee as long as the protectors of data sets that share subjects coordinate or follow suitable strategies (Domingo-Ferrer and Soria-Comas, 2016). On the other hand, there are heuristics that allow reaching $k$-anonymity at quasi-linear cost in the number of records. At the same time, it is possible to link the information of the same subject in several independently $k$-anonymized data sets, at least at the level of a $k$-anonymous class (and in some cases at the record level). Also, utility improves when parameter $k$ is smaller and/or the set of subjects is larger, because less modification of the original records is needed to reach $k$-anonymity. Nonetheless, reducing $k$ also reduces protection. *All in all, with some coordination effort, k-anonymity can be a starting point to anonymize big data.*

# 4 BIG DATA PROTECTION UNDER DIFFERENTIAL PRIVACY

A randomized query function $F$ gives $\varepsilon$-differential privacy (DP) if, for all data sets $D_1$, $D_2$ such that one can be obtained from the other by modifying a single record (neighboring data sets), and all $S \subset Range(F)$,

$$\Pr(F(D_1) \in S) \leq \exp(\varepsilon) \times \Pr(F(D_2) \in S). \quad (1)$$

A common SDC method to enforce DP is Laplace noise addition. Although the original definition above was for interactive queries to databases, DP was later extended for anonymizing data sets (Soria-Comas *et al.*, 2014), (Xiao *et al.*, 2010), (Xu *et al.*, 2012), (Zhang *et al.*, 2014).

*Protection* under DP is very good if $\varepsilon$ is small, because in that case Expression (1) ensures that the presence or absence of any single individual cannot be noticed from the anonymized output.

The biggest problem of DP is that it offers very poor *utility* for exploratory analyses fot the values of $\varepsilon$ that are required to ensure a good privacy (typically below 1, see Section 5 below). Indeed, enough noise must be added to make the presence or absence of any individual unnoticeable, and this includes *any* outlying individual that can potentially be part of the data set. Thus, for small $\varepsilon$, the noise to be added is huge.

Regarding *composability*, DP offers strong properties (McSherry, 2009):

- *Sequential composition*. If, for the same subset of subjects, data sets $D_i$ are released with $i \in I$, each of which is $\varepsilon_i$-diferentially private, the pooled released data sets offer $\sum_{i \in I} \varepsilon_i$-differential privacy. That is, when several differentially private data sets on a set of subjects are collated, differential privacy is not broken, but the level of privacy is reduced.
- *Parallel composition*. If several $\varepsilon$-differentially private data sets $D_i$ are released, for $i \in I$, *each one referring to a disjoint set of subjects*, all those data sets taken together are $\varepsilon$-differentially private.

As to *computational cost*, DP is attained by adding noise to original data. The cost of adding noise is linear in the number $n$ of records, that is, $O(n)$.

Finally, in what concerns *linkability*, if we have two differentially private data sets in which noise has been added to all values of all attributes, in general it is not possible to link the records in both data sets that correspond to the same subject. However, if both data sets share some attributes that have been left unmodified (e.g. because they are not deemed confidential), then the values of those attributes can be used to link the records corresponding to the same subject.

Thus, differential privacy has some interesting properties for big data: good composability, good computational cost, and linkability if there are shared and unmodified attributes across several data sets. The killing problem is the lack of utility of DP data, especially for data uses that could not be anticipated by the data protector.

Some authors have suggested computational procedures different from mere noise addition to improve the utility of differentially private data (Cormode *et al.*, 2012), (Zhang *et al.*, 2014), (Sánchez *et al.*, 2014), (Sánchez *et al.*, 2016b). Unfortunately, these more sophisticated procedures are computationally more demanding than simple noise addition and they provide rather small utility increases, insufficient for exploratory analyses of reasonable quality. The reason is that, as explained above, the very definition of differential privacy requires destroying the information of each record so that its presence or absence cannot be noticed.

# 5 CONNECTIONS BETWEEN PRIVACY MODELS

Since none of the two big families of privacy models is entirely satisfactory for big data, let us examine their connections and underlying principles. That may shed light on possible ways to adapt or replace them.

We will show that the following privacy models are interconnected around the principles of *deniability* and *permutation*: randomized response (RR), post-randomization (PRAM), differential privacy (DP) and $t$-closeness. Specifically, we will characterize RR in terms of deniability and then we will show that PRAM can be view as RR done by the data controller. Then we will use the connection between RR and DP to explain the effect of taking large $\varepsilon$ in DP in terms of deniability. After that, we will use the connection between DP and $t$-closeness to explain DP in terms of the intruder's knowledge gain on the sensitive attribute. Finally, we will present PRAM in terms of the permutation paradigm; since PRAM is connected to RR, and RR to DP and $t$-closeness, it will turn out that all those models can in fact be viewed as permutation. This is a novel result.

## 5.1 Randomized Response, Plausible Deniability and PRAM

Let $X$ be an attribute containing the answer to a sensitive question. If $X$ can take $r$ possible values, then the randomized response $Y$ (Greenberg *et al.*, 1969) reported by the respondent instead of $X$ is computed using

$$\mathbf{P} = \begin{pmatrix} p_{11} & \cdots & p_{1r} \\ \vdots & \vdots & \vdots \\ p_{r1} & \cdots & p_{rr} \end{pmatrix}$$

where $p_{uv} = \Pr(Y = v | X = u)$, for $u, v \in \{1, \ldots, r\}$ denotes the probability that the randomized response is $v$ when the respondent's true attribute value is $u$.

Let $\pi_1, \ldots, \pi_r$ be the proportions of respondents whose true values fall in each of the $r$ categories of $X$; let $\lambda_v = \sum_{u=1}^{r} p_{uv}\pi_u$ for $v = 1, \ldots, r$, be the probability of the reported value $Y$ being $v$. If we define $\lambda = (\lambda_1, \ldots, \lambda_r)^T$ and $\pi = (\pi_1, \ldots, \pi_r)^T$, then $\lambda = \mathbf{P}^T \pi$. Furthermore, if $\hat{\lambda}$ is the vector of sample proportions corresponding to $\lambda$ and $\mathbf{P}$ is nonsingular, it is proven in (Chaudhuri and Mukherjee, 1988) that an unbiased estimator of $\pi$ can be obtained as

$$\hat{\pi} = (\mathbf{P}^T)^{-1}\hat{\lambda}.$$

**Privacy Guarantees of RR.** The privacy guarantees RR offers to respondents are *plausible deniability* and *secrecy*:

- By the Bayes' formula:

$$\hat{p}_{vu} = \Pr(X = u | Y = v) = \frac{p_{uv}\pi_u}{\sum_{u'=1} p_{u'v}\pi_{u'}}.$$

- Given a reported $Y = v$, *deniability* can be measured as

$$H(X|Y=v) = -\sum_{u=1}^{r} \hat{p}_{vu} \log_2 \hat{p}_{vu}.$$

- If the probabilities within each column of $\mathbf{P}$ are identical, then $\hat{p}_{vu} = \pi_u$, for $u, v \in \{1, \ldots, r\}$, and $H(X|Y=v) = H(X)$ for any $v$, and thus $H(X|Y) = H(X)$ (*Shannon's perfect secrecy*).
- The price paid for perfect secrecy is a singular matrix $\mathbf{P}$, so no unbiased estimator $\hat{\pi}$ can be computed.

**Randomized Response: a Local Version of PRAM.** Matrix $\mathbf{P}$ looks exactly as the PRAM transition matrix (Gouweleeuw *et al.*, 1998). The main difference is that in RR randomization is done by the respondent, whereas in PRAM it is done by the data controller. Thus, *RR is a local anonymization method avant la lettre*: when RR was invented, the notion of anonymization did not exist, let alone local anonymization.

## 5.2 Randomized Response and Differential Privacy

(Wang *et al.*, 2016) show that RR is ε-differentially private if

$$e^{\varepsilon} \geq \max_{v=1,\ldots,r} \frac{\max_{u=1,\ldots,r} p_{uv}}{\min_{u=1,\ldots,r} p_{uv}}. \quad (2)$$

We can assert that:

- If the maximum ratio between the probabilities in a column of $\mathbf{P}$ is bounded by $e^{\varepsilon}$, the influence of the real value $X$ on the reported value $Y$ is limited.
- When $\varepsilon = 0$, in bound (2), the probabilities within each column of $\mathbf{P}$ are identical, and RR provides perfect secrecy. Thus, *DP with strictest privacy ($\varepsilon = 0$) offers perfect secrecy*.

**Explaining Large ε in DP using Deniability.** When one takes not-so-small ε, the intuition of DP is unclear: it is no longer tenable that the presence or absence of any single record is unnoticeable. *The connection of DP with RR and hence with deniability helps understand what large ε implies.*

**Example 1.** if $\varepsilon = 2$, in some columns of $\mathbf{P}$ the probability ratio may be as large as $e^2 = 7.389$. If $r = 2$, one might have a column with $p_{1v} = 0.7389$ and $p_{2v} = 0.1$. Thus, after reporting $Y = v$, the most likely value is $X = 1$ and there is only a small margin to deny it. *Thus, clearly $\varepsilon = 2$ does not seem to offer enough privacy.*

## 5.3 Differential Privacy and $t$-closeness

A data set is said to satisfy $t$-closeness if, for each group of records sharing a combination of quasi-identifier attributes, the distance between the distribution of the confidential attribute in the group and the distribution of the attribute in the whole data set is no more than a threshold $t$.

Given two random distributions $F_1$ and $F_2$, consider the distance

$$d(F_1, F_2) = \max_{i=1,2,\cdots,t} \left\{ \frac{\Pr_{F_1}(x_i)}{\Pr_{F_2}(x_i)}, \frac{\Pr_{F_2}(x_i)}{\Pr_{F_1}(x_i)} \right\}. \quad (3)$$

In Expression (3), we take the quotients of probabilities to be zero if both $\Pr_{F_1}(x_i)$ and $\Pr_{F_2}(x_i)$ are zero, and to be infinity if only the denominator is zero. In (Domingo-Ferrer and Soria-Comas, 2015), we showed the following connection between ε-DP and $t$-closeness:

**Proposition 1.** *Let $k_I(D)$ be the function that returns the view on subject I's sensitive attributes given a data set D. If D satisfies $\exp(\varepsilon/2)$-closeness when using the distance in Expression (3), then $k_I(D)$ satisfies ε-differential privacy. In other words, if we restrict the domain of $k_I$ to $\exp(\varepsilon/2)$-close data sets, then we have ε-differential privacy for $k_I$.*

Proposition 1 can explain DP in terms of the intruder's knowledge gain on the sensitive attribute value of a target respondent if the intruder can determine the respondent's cluster.

**Example 2.** Take DP with $\varepsilon = 2$. By the proposition, the probability weight attached to a certain value of a sensitive attribute $X$ can grow by a factor $e \approx 2.718$ if the target individual's cluster is learnt by the intruder.

To decide whether a probability has grown too much, consider that the reported value $v$ is the cluster identifier and probabilities $\hat{p}_{vu} = \Pr(X = u|Y = v)$, for $u = 1, \ldots, r$ are the probabilities assigned by the cluster-level distribution to the values of the sensitive attribute within the cluster. Determining the real $X$ given the reported $Y$ becomes determining the target respondent's sensitive value $X$ given the target respondent's cluster $Y$. We can use a deniability argument to assess whether the cluster-level distribution is too inhomogeneous:

**Example 3.** Take $\varepsilon = 2$ and assume the sensitive attribute can take $r = 5$ different values, with uniform data set-level distribution (prob. $1/5$ for each value). A cluster-level distribution with one value having relative frequency $1/5 \times \exp(1) = 0.5436$ and the remaining four values $0.1141$ satisfies $\exp(1) -$ *closeness*. The cluster-level distribution makes guessing the sensitive attribute value much easier than the

data set-level distribution (thus $\varepsilon = 2$ does not offer enough protection).

In Example 1, we used the connection between differential privacy and randomized response to illustrate the weak privacy protection by differential privacy when $\varepsilon$ is not sufficiently small. Example 3 provides yet another way to assess the effects of taking large $\varepsilon$ in differential privacy, this time based on the connection with $t$-closeness.

### 5.4 PRAM and the Permutation Paradigm

**Reverse Mapping.** In (Domingo-Ferrer and Muralidhar, 2016), the following procedure was described:

**Require:** Original attribute $X = \{x_1, x_2, \cdots, x_n\}$
**Require:** Anonymized attribute $Y = \{y_1, y_2, \cdots, y_n\}$
  **for** $i = 1$ to $n$ **do**
    Compute $j = \text{Rank}(y_i)$
    Set $z_i = x_{(j)}$ (where $x_{(j)}$ is the value of $X$ of rank $j$)
  **end for**
  **return** $Z = \{z_1, z_2, \cdots, z_n\}$

**The Permutation Paradigm.** The output $Z$ is a permutation of $X$ and has the same rank order as $Y$. Thus any anonymization procedure can be viewed as a permutation ($X$ into $Z$) followed by residual noise addition ($Z$ into $Y$) that does not alter ranks.

**PRAM and the Permutation Paradigm.** PRAM does not permute attribute values in the data set, rather it permutes in the *domain* of attributes. Hence, *PRAM should be viewed in terms of the permutation paradigm as permutation plus noise.* Hence, *RR can also be viewed as permutation, and so can DP and $t$-closeness.*

## 6 CONCLUSIONS AND FURTHER RESEARCH

There is a debate on whether big data are compatible with the privacy of citizens. We have stated the desirable properties of privacy models for big data (protection, utility, composability, low computation and linkability). We have examined how well the two main privacy models ($k$-anonymity and $\varepsilon$-differential privacy) satisfy those properties. None of them is entirely satisfactory, although $k$-anonymity seems more amenable to big data anonymization. We have highlighted connections between the main privacy models that might result in synergies between them in order to tackle big data: the principles underlying all those models are deniability and permutation.

Future research will have to deal with adapting the current privacy models for big data and/or designing new models. When tackling this endeavor, it will essential to adhere to the above-mentioned principles of deniability and permutation.

## ACKNOWLEDGMENTS AND DISCLAIMER

Partial support for this work has been received from the Government of Catalonia (ICREA Acadèmia Award), the European Commission (projects H2020-644024 “CLARUS” and H2020-700540 “CANVAS”) and the Spanish Government (project TIN2014-57364-C2-1-R “SmartGlacis”). I hold the UNESCO Chair in Data Privacy, but the opinions in this paper are my own and do not commit UNESCO.

## REFERENCES

Chaudhuri, A., and Mukherjee, R. (1988) *Randomized Response: Theory and Techniques*. Marcel Dekker.

Cormode, G., Procopiuc, C., Srivastava, D., Shen, E., and Yu, T. (2012) Differentially private spatial decompositions. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering-ICDE'12*, Washington, DC, USA, pp. 20-31. IEEE Computer Society.

Domingo-Ferrer, J. (2007) A three-dimensional conceptual framework for database privacy. In *4th VLDB Workshop on Secure Data Management-SDM'07*, pp. 193-202. Springer.

Domingo-Ferrer, J., and Muralidhar, K. (2016) New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. *Information Sciences*, 337-338:11-24.

Domingo-Ferrer, J., and Soria-Comas, J. (2015) From $t$-closeness to differential privacy and vice versa in data anonymization. *Knowledge-Based Systems*, 74:151-158.

Domingo-Ferrer, J., and Soria-Comas, J. (2016) Anonymization in the time of big data. In *Privacy in Statistical Databases-PSD 2016*, pp. 225-236. Springer.

Domingo-Ferrer, J., and Torra, V. (2005) Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining & Knowledge Discovery*, 11(2):195-212.

Dwork, C. (2006) Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming-ICALP'06*, pp. 1-12. Springer.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L.C.R.J, and De Wolf, P.-P. (1998) Post randomisation for statistical disclosure control: theory and implementation. *Journal of Official Statistics*, 14:463-478.

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., and Horvitz, D. G. (1969) The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, 64(326):520-539.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., and De Wolf, P.-P. (2012) *Statistical Disclosure Control*. Wiley.

Li, N., Li, T., and Venkatasubramanian, S. (2007) t-Closeness: privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd IEEE International Conference on Data Engineering-ICDE 2007*, Istanbul, Turkey, pp. 106-115. IEEE Computer Society.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007) l-Diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1).

McSherry, F. D. (2009) Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data-SIGMOD'09*, New York, NY, USA, pp. 19-30. ACM.

Samarati, P., and Sweeney, L. (1998) *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. Technical Report, SRI International.

Sánchez, D., Domingo-Ferrer, J., and Martínez, S. (2014) Improving the utility of differential privacy via univariate microaggregation. In *Privacy in Statistical Databases-PSD 2014*, pp. 130-142. Springer.

Sánchez, D., Domingo-Ferrer, J., Martínez, S., and Soria-Comas, J. (2016) Utility-preserving differentially private data releases via individual ranking microaggregation. *Information Fusion*, 30:1-14.

Sánchez, D., Martínez, S., and Domingo-Ferrer, J. (2016b) Comment on 'Unique in the shopping mall: on the reidentifiability of credit card metadata'. *Science*, 351(6279), pp. 1274. March 18.

Soria-Comas, J., and Domingo-Ferrer, J. (2012) Probabilistic k-anonymity through microaggregation and data swapping. In *Proceedings of the IEEE International Conference on Fuzzy Systems-FUZZ-IEEE 2012*, Brisbane, Australia, pp. 1-8. IEEE.

Soria-Comas, J., and Domingo-Ferrer, J. (2015) Big data privacy: challenges to privacy principles and models. *Data Science and Enginering,* 1(1):21-28.

Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., and Martínez, S. (2014) Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *VLDB Journal* 23(5):771-794.

U. S. Federal Trade Commission (2014) *Data Brokers: A Call for Transparency and Accountability*.

Wang, Y., Xu, X., and Hu, D. (2016) Using randomized response for differential privacy preserving data collection. In *EDBT/ICDT 2016 Joint Conference*, Bordeaux, France.

Xiao, Y., Xiong, L., and Yuan, C. (2010) Differentially private data release through multidimensional partitioning. In *Proceedings of the 7th VLDB Conference on Secure Data Management - SDM'10*, pp. 150-168. Springer.

Xu, J., Zhang, Z., Xiao, X., Yang, Y., and Yu, G. (2012) Differentially private histogram publication. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering-ICDE'12*, Washington, DC, USA, pp. 32-43. IEEE Computer Society.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. (2014) Privbayes: private data release via Bayesian networks. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data-SIGMOD'14*, New York, NY, USA, pp. 1423-1434. ACM.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS