Data Sci. Eng. (2016) 1(1):21–28
DOI 10.1007/s41019-015-0001-x

INVITED PAPER

Big Data Privacy: Challenges to Privacy Principles and Models

Jordi Soria-Comas[1] · **Josep Domingo-Ferrer**[1]

Received: 12 August 2015 / Revised: 12 August 2015 / Accepted: 16 August 2015 / Published online: 15 September 2015

Abstract This paper explores the challenges raised by big data in privacy-preserving data management. First, we examine the conflicts raised by big data with respect to preexisting concepts of private data management, such as consent, purpose limitation, transparency and individual rights of access, rectification and erasure. Anonymization appears as the best tool to mitigate such conflicts, and it is best implemented by adhering to a privacy model with precise privacy guarantees. For this reason, we evaluate how well the two main privacy models used in anonymization (k-anonymity and ε-differential privacy) meet the requirements of big data, namely composability, low computational cost and linkability.

Keywords Big data · Consent · Privacy models · k-anonymity · ε-differential privacy

1 Introduction

Big data have become a reality in recent years: Data are being collected by a multitude of independent sources, and then they are fused and analyzed to generate knowledge. Big data depart from previous data sets in several aspects such as volume, variety and velocity. The large amount of data has put too much pressure on traditional structured data stores, and as result new technologies have appeared, such as Hadoop, NoSQL, MapReduce [7]. The amount and variety of data have made sophisticated data analyses possible. Data analysis is no longer only a matter of describing data or testing hypotheses, but also of generating (previously unavailable) knowledge out of the data.

While a valuable resource in many fields, there is an important side effect of big data. The privacy of the individuals whose data are being collected and analyzed (often without their being aware of it) is increasingly at risk. An illustrative case of this is reported in [9]. Target, a large retailer, created a model for pregnancy prediction. The goal was to send discount coupons on several baby-related products as soon as possible with the aim to shape long-standing buying patterns to Target's advantage. Some time later, a father complained to Target that her daughter, still at high school, had been sent coupons for baby clothes; he asked whether they were encouraging her to get pregnant. It was later discovered that she was pregnant but her father was still unaware of it.

While in a different setting and scale, disclosure risk has long been a concern in the statistical and computer science communities, and several techniques for limiting such risk have been proposed. Statistical disclosure control (SDC, [14]) seeks to allow one to make useful inferences about subpopulations from a data set while at the same time preserving the privacy of the individuals that contributed their data. Several SDC techniques have been proposed to limit the disclosure risk in microdata releases. A common feature in all of them is that the original data set is kept secret and only a modified (anonymized) version of it is released. In recent years, several privacy models have been proposed. Rather than determining the specific transformation that should be carried out on the original data, privacy models specify conditions that the data set must satisfy to keep disclosure risk under control. Privacy models usually depend on one or several parameters that determine how

✉ Josep Domingo-Ferrer
josep.domingo@urv.cat

Jordi Soria-Comas
jordi.soria@urv.cat

[1] Department of Computer Engineering and Mathematics, UNESCO Chair in Data Privacy, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Catalonia

much disclosure risk is acceptable. Existing privacy models have been mostly developed with a single static data set in mind. However, with big data this setting does not suffice any more.

We next sketch the contributions and the structure of this paper. In Sect. 2, we examine the conflict between big data and the legal and ethical requirements in private data management. Given that anonymization appears as the best option to mitigate that conflict, and since privacy models seem the soundest approach to anonymization, in Sect. 3 we seek to determine the properties that a privacy model must have to be appropriate for big data privacy. In Sects. 4 and 5, respectively, we evaluate the two best-known privacy models, namely k-anonymity [24] and differential privacy [11], in terms of the desirable properties for big data protection established in Sect. 3. Finally, Sect. 6 summarizes the conclusions of our work.

2 Privacy in Big Data

The potential risk to privacy is one of the greatest downsides of big data. It should be taken into account that big data are all about gathering as many data as possible to extract knowledge from them (possibly in some innovative ways). Moreover, more than often, these data are not consciously supplied by the data subject (typically a consumer, citizen), but they are generated as a by-product of some transaction (e.g., browsing or purchasing items in an online store), or they are obtained by the service provider in return for some free service (for example, free email accounts, social networks) or as a natural requirement for some service (e.g., a GPS navigation system needs knowledge about the position of an individual to supply her with information on nearby traffic conditions).

At the moment, there is not a clear view of the best strategy or strategies to protect privacy in big data. Prior to the advent of big data, the following principles were of broad application in several regulations for the protection of personally identifiable information (PII) [6]:

- *Lawfulness*. Consent needs to be obtained from the subject, or the processing must be needed for a contract or legal obligation, for the subject's vital interests, for a public interest, or for legitimate processor's interests compatible with the subject's rights.
- *Consent*. The consent given by the subject must be simple, specific, informed and explicit.
- *Purpose limitation*. The purpose of the data collection must be legitimate and specified before the collection.
- *Necessity and data minimization*. Only the data needed for the specific purpose should be collected. Furthermore, the data must be kept only for as long as necessary.
- *Transparency and openness*. Subjects need to get information about data collection and processing in a way they can understand.
- *Individual rights*. Subjects should be given access to the data on them, as well as the possibility to rectify or even erase such data (right to be forgotten).
- *Information security*. The collected data must be protected against unauthorized access, processing, manipulation, loss or destruction.
- *Accountability*. The data collector/processor must have the ability to demonstrate compliance with the above principles.
- *Data protection by design and by default* [6]. Privacy must be built-in from the start rather than added later.

Without anonymization, several potential conflicts appear between the above principles and the purpose of big data:

- *Purpose limitation*. Big data are often used secondarily for purposes not even known at the time of collection.
- *Consent*. If the purpose of data collection is not clear, consent cannot be obtained.
- *Lawfulness*. Without purpose limitation and consent, lawfulness is dubious.
- *Necessity and data minimization*. Big data result precisely from accumulating data for potential use.
- *Individual rights*. Individual subjects do not even know which data are stored on them or even who holds data on them. Accessing, rectifying or erasing the data is therefore infeasible for them.
- *Accountability*. Compliance does not hold and hence it cannot be demonstrated.

Given the above conflicts between privacy principles and big data, it has been argued that, in order to avoid hampering technological development, privacy protection should focus only on potentially privacy-harming uses of the data (rather than on data collection) or even allow for self-regulation. In the opposite camp, it has been also argued that it is the mere collection of the data that triggers any potential privacy breaches. Indeed, once the data have been collected, many potential threats arise [3]:

- *Data breach.* This may happen as a result of aggressive hacking or insufficient security measures. The more data are collected, the more appealing they become for an attacker.
- *Internal misuse by employees* [4].
- *Unwanted secondary use*.
- *Changes in company practices*. The policies that prevent a company from engaging in uses of the data that harm the subjects' interests may change.
- *Government access without due legal guarantees* [27].

Anonymization techniques are a possible solution to overcome the conflicts between privacy principles and big data. As the privacy principles above refer to PII, once the data have been anonymized they may be viewed as being no longer PII and hence one may claim that principles no longer apply to them. However, anonymization techniques face some difficulties when applied to big data. On one side, too little anonymization (e.g., mere de-identification by just suppressing direct identifiers) may not be enough to ensure non-identifiability [2,13]. This becomes more problematic with big data because, as the amount and variety of data about an individual accumulates, re-identification becomes more plausible. On the other side, too strong an anonymization may prevent linking data on the same individual subject (or on similar individual subjects) that come from different sources and, thus, thwart many of the potential benefits of big data.

While it is obvious that there are some tensions between big data and data anonymization, we should not rule out the latter. Admittedly, it is true that anonymization can hamper some of the uses of big data (mainly those uses targeting a specific individual), but anonymized data still enable most analyses in which the target is a sufficiently large community or the entire population.

3 Properties of Privacy Models for Big Data

SDC techniques [14] (e.g., global recoding, supression, top and bottom coding, microaggregation) specify data transformations whose purpose is to limit disclosure risk. Yet, in general they do not specify any mechanism to assess what is the disclosure risk remaining in the transformed data. On the other side, privacy models (such as k-anonymity [24], l-diversity [18], t-closeness [17], ε-differential privacy [11], probabilistic k-anonymity [28]) specify some properties that the data set must satisfy to limit disclosure risk, but they leave it open which SDC technique is to be used to satisfy these properties. In this sense, privacy models seem more appealing. The reality, however, is that most privacy models have been designed to protect a single static original data set and, thus, there are several important limitations in their application to big data settings.

The three characteristics often mentioned as distinctive of big data are volume, variety and velocity. Volume refers to the fact that the amount of data is are subject to analysis is large. Variety refers to the fact that big data consist of heterogeneous types of data extracted and fused from several different sources. Velocity refers to the speed of generation and the processing of the data. Certainly, not all the above properties need to concur for the name big data to be used, but at least some of them are required. Even though volume is reflected in the name of "big" data, usually variety is regarded as the most relevant feature of big data.

For a privacy model to be usable in a big data environment, it must cope well with volume, variety and velocity. To determine the suitability of a privacy model for big data, we look at the extent to which it satisfies the following three properties:

- *Composability*. A privacy model is composable if the privacy guarantees of the model are preserved (possibly to a limited extent) after repeated independent application of the privacy model. From the opposite perspective, a privacy model is not composable if multiple independently data releases, each of them satisfying the requirements of the privacy model, may result in a privacy breach.
- *Computational cost*. The computational cost measures the amount of work needed to transform the original data set into a data set that satisfies the requirements of the privacy model. We have previously mentioned that, usually, there is a variety of SDC techniques that can be employed to satisfy the requirements of the privacy model. Thus, the computational cost depends on the particular SDC technique selected. When evaluating the cost of a privacy model, we will consider the most common approaches.
- *Linkability*. In big data, information about an individual is gathered from several independent sources. Hence, the ability to link records that belong to the same (or a similar) individual is central in big data creation. With privacy protection in mind, the data collected by a given source should be anonymized before being released. However, this independent anonymization may limit the data fusion capabilities, thereby severely restricting the range of analyses that can be performed on the data and, consequently, the knowledge that can be generated from them. The amount of linkability compatible with a privacy model determines whether and how an analyst can link data independently anonymized under that model that correspond to the same individual. Notice that, when linking records belonging to the same individual, we are increasing the information about this individual. This is a privacy threat, and thus, the accuracy of the linkages should be less in anonymized data sets than in original data sets.

All of the above properties of a privacy model seem to be important to effectively deal with big data. We next discuss the importance of each property in some detail.

Composability is essential for the privacy guarantees of a model to be meaningful in the big data context. In big data, the process of data collection is not centralized, but distributed among several data sources. If one of the data collectors is concerned about privacy and decides to use a specific privacy model, the privacy guarantees of the selected model should be preserved (to some extent) after the fusion of the data. Composability can be evaluated between data releases that satisfy

the same privacy model, between data releases that satisfy different privacy models and between a data release that satisfies a privacy model and non-anonymized data. In this paper, we evaluate only composability between data releases that satisfy the same privacy model. The strongest case, composability against a non-anonymized data release, is equivalent to requiring the privacy model to offer protection against arbitrary side knowledge.

Low computational cost is also a very important feature of a privacy model if it has to remain practical for big data, given that one of the main properties of big data is volume (that is, large data sets). Algorithms having linear or log-linear cost on the size of the data set seem to be feasible alternatives. Algorithms with quadratic cost or above are not feasible for large data sets. Several simple modifications on a costly algorithm are conceivable to make it more efficient. One such modification is to partition the original data set in several smaller data sets and anonymize each of these separately. Of course, partitioning may have implications on the utility and the privacy of the resultant data; such implications ought to be analyzed on a case-by-case basis.

Finally, linkability is also essential as big data are obtained by fusing inputs from different sources. More precisely, being able to link data about the same individual in different data sets is of foremost importance for data analysis. Thus, to be useful for big data, a privacy model must allow anonymized data to stay linkable to some extent.

4 Evaluating *k*-Anonymity

k-Anonymity seeks to limit the disclosure risk of a data set by limiting the capability of intruders to re-identify a record, that is, *k*-anonymity seeks to prevent *identity disclosure*. To that end, *k*-anonymity assumes that record re-identification is performed based on a fixed set of combinations of attributes. Each such combination is known as a *quasi-identifier*. The goal in *k*-anonymity is to make the combination of values of a quasi-identifier in the anonymized data set to refer to at least *k* individuals; these individuals form a so-called equivalence class.

Definition 1 (*k-anonymity* [24]) Let $T[A_1, \ldots, A_n]$ be a microdata set with attributes A_1 to A_n, and let QI be a quasi-identifier associated with it. T is said to satisfy *k*-anonymity with respect to QI if and only if each sequence of values in $T[QI]$, the projection of T over the attributes in QI, appears at least k times in $T[QI]$.

A quasi-identifier is a set of attributes in the data set that are externally available in combination (i.e., appearing together in an external data set or in possible joins between external data sets) and associated with an identified individual. However, determining which combinations of attributes should be taken as quasi-identifiers is controversial, as it seems difficult to argue that the data protector knows all such combinations of attributes. If we are to protect against informed intruders (with knowledge of some confidential information), probably all attributes in the data set should be taken as a quasi-identifier.

Algorithm 1 Intersection attack

$R_1, \ldots, R_n \leftarrow n$ independent data releases
$P \leftarrow$ population consisting of subjects present in all $R_1, \ldots, R_n$
for each individual i **in** P **do**
  for $j = 1$ **to** n **do**
    $e_{ij} \leftarrow$equivalence class of R_j associated to i
    $s_{ij} \leftarrow$set of sensitive values of e_{ij}
  end for
  $S_i \leftarrow s_{i1} \cap s_{i2} \cap \ldots \cap s_{in}$
end for
return $S_1, \ldots, S_{|P|}$

It is important to keep in mind that disclosure may occur even without the re-identification of a record. Attribute disclosure happens when access to the released data gives the intruder a better knowledge about the value of a confidential attribute of a specific individual. In a *k*-anonymous data set, there is attribute disclosure if the variability of the confidential attributes in a *k*-anonymous group of records is small. Several extended models have been proposed that try to address this shortcoming of *k*-anonymity; *l*-diversity [18] and *t*-closeness [17] are among them. We do not analyze these extended models in this paper; however, most of the big data related properties of *k*-anonymity also apply to such extensions.

4.1 Composability

k-Anonymity has been designed to limit the disclosure risk in a single static data set; thus, in general, the *k*-anonymity model does not compose. A good analysis of the composability properties of *k*-anonymity can be found in [12]. To break the guarantees of *k*-anonymity in multiple data releases, the intruder tries to use the information about the confidential attribute provided by each of the data releases to further restrict the feasible values. This is the way that the *intersection attack* [12] proceeds (see Algorithm 1). This attack can lead to reduced privacy guarantees (a smaller k) or even to exact re-identification of some records.

In spite of *k*-anonymous data releases being in general not composable, in some restricted scenarios composability may be satisfied. By observing Algorithm 1, it is clear that two conditions must be satisfied for the attack to be succesful: (i) There must be some overlapping subjects across the independent data releases, and (ii) information about the same confidential attribute must be present in several data releases. If we can guarantee that there are no overlapping subjects or confidential attributes in the *k*-anonymous data

releases, k-anonymity is preserved. However, it seems difficult to guarantee such conditions in a big data environment.

4.2 Computational Cost

Attaining k-anonymity requires modifying the original data in such a way that the combination of values of each quasi-identifier is shared by at least k records. Although this could be achieved in many different ways, in order to minimize the information loss one usually resorts to clustering quasi-identifier values that are as close as possible (given some previously defined closeness criterion).

Generalization (*a.k.a.* recoding) and suppression are two commonly used techniques to attain k-anonymity. Initial approaches were based on one-dimensional global recoding, where a hierarchy of generalizations is defined for each individual attribute and the goal is to find a minimal generalization over all the attributes that satisfies k-anonymity. Finding the optimal generalization based on these primitives has been shown to be an NP-hard problem [1,21]. Several cost-effective heuristics for finding an approximate solution have been proposed. In [24], an algorithm based on binary search among the set of possible generalizations is proposed for a specific notion of minimality. In [15], an algorithm is proposed that is capable of dealing with an arbitrary minimality notion. Although both [24] and [15] have, in the worst case, an exponential cost on the size of the quasi-identifier, the optimizations they perform allow them to deal with relatively large quasi-identifiers [15]. The cost is linear in the size of the data set. Multi-dimensional global recoding can be used to attain k-anonymity, instead of one-dimensional global recoding. The additional flexibility of multi-dimensional global recoding allows for improved utility. As for the running cost, optimal multi-dimensional global recoding is an NP-hard problem, but an approximation algorithm with cost $\mathcal{O}(n \ln n)$ in the size of the data set has been proposed in [16].

Microaggregation can be used as an alternative to global recoding to attain k-anonymity. The optimal microaggregation has been shown to be an NP-hard problem [23], but several approximation heuristics have been proposed that reduce its cost. For instance, MDAV [8] has a $\mathcal{O}(n^2)$ cost in the size n of the data set. Partitioning the data set and separately microaggregating each subset resulting from the partition is a way to make MDAV usable on large data sets.

Which of the above approaches is the best performer depends on the actual data set (see Table 1). For one-dimensional global recoding the cost is exponential in the size of the quasi-identifier, so it is not a valid option beyond a given size of the QI. In spite of this, its cost is linear in the number of records; thus, it is a good option for dealing with large data sets with quasi-identifiers consisting of relatively few attributes. See [15] for empirical evaluations. When one-dimensional global recoding is not applicable because of the size of the quasi-identifier, we can resort to multi-dimensional global recoding and microaggregation. Multi-dimensional global recoding, with cost $\mathcal{O}(n \ln n)$ in the size of the data set, seems a feasible approach even for the large values of n that can be expected in a big data environment. On the other hand, although microaggregation has $O(n^2)$ cost, it has appealing utility features, especially for numerical attributes (see [8]), and it can still be used for large n if combined with partitioning.

Table 1 Cost of attaining k-anonymity in terms of the primitives used

Primitive	Cost	Algorithm
Single-dim. global recoding	Exp. in the size of the QI linear in the no. of records	Incognito
Multi-dim. global recoding	Quasilinear in the no. of records	Mondrian
Microaggregation	Quadratic in the no. of records	MDAV

4.3 Linkability

The scenario we consider for evaluating linkability consists of two independently anonymized data sets whose populations of subjects partially overlap. We want to determine whether it is possible to link the records that belong to the same individual.

For a subject that is known to be in both data sets, we can determine the corresponding groups of k-anonymous records containing her (approximately if microaggregation is used instead of generalization). Thus, we can at least link the groups of k-anonymous records containing the given individual. If some of the confidential attributes are shared between the data sets, the accuracy of the linkage improves. It could even be possible to accurately link individual records (not just k-anonymous groups).

5 Evaluating Differential Privacy

Differential privacy [10,11] is a privacy model offering strong privacy guarantees. It seeks to limit the impact of any single individual subject's contribution on the outcome of any analysis. Its primary setting is based on a trusted party that holds the original data set, receives queries and returns randomized answers for those queries so that the following differential privacy guarantee is satisfied.

Definition 2 A randomized function κ gives ε-differential privacy if, for all data sets D_1 and D_2 that differ in one record, and all $S \subset Range(\kappa)$

$$\Pr(\kappa(D_1) \in S) \leq \exp(\varepsilon) \times \Pr(\kappa(D_2) \in S).$$

In the definition above, κ represents the (randomized) query/analysis that the data user wants to perform on the data. Because the data sets D_1 and D_2 differ in one record (where each record corresponds to a subject), and differential privacy seeks to limit the impact of each single subject in the result of any analysis, the outcomes $\kappa(D_1)$ and $\kappa(D_2)$ must be similar. This similarity is measured by differential privacy in terms of the certainty of getting a specific outcome. The level of privacy is controlled by the parameter ε. The smaller ε, the more privacy, but on the other hand, more noise needs to be introduced to randomize the query outcome (and more noise means less utility).

Several approaches have been proposed to generate differentially private data sets [19,29–32], although the main purpose of differential privacy remains to provide privacy-preserving answers for queries performed on the original data. However, data analysts usually expect to have access to the underlying data set, rather than be given noise-added answers to specific interactive queries. For this reason, this paper focuses on differentially private data sets.

There are two main approaches to generate differentially private data sets: (i) create a synthetic data set from an ε-differentially private model for the data (usually from a differentially private histogram), or (ii) add noise to mask the values of the original records (probably in combination with some prior aggregation function to reduce the amount of required noise).

A synthetic data set can be either partially synthetic or fully synthetic. In partially synthetic data sets, only some of the values of the original data set are replaced by synthetic (simulated) values, usually the ones that are deemed too sensitive. On the contrary, in fully synthetic data sets a new data set is generated from scratch. When generating a fully synthetic data set, the original data set is viewed as a sample from some underlying population and the synthetic data set is generated by taking a different sample. The attributes in the original data set are split in two groups (see Fig. 1): those that are known for the entire population (labeled A in the figure), and those that are known only for the sample (labeled B in the figure). A_{pop} refers to the matrix of values that are available for all population subjects (in a survey, these are usually the attributes used in the design of the sample), B_{obs} refers to the part of the matrix B with available (observed) values and B_{mis} refers to the part with missing (unobserved) values. To generate the fully synthetic data set, the values of B_{mis} are imputed (by adjusting a model for the joint distribution of (A, B) to the observed data and then drawing from it conditional to the value of A).

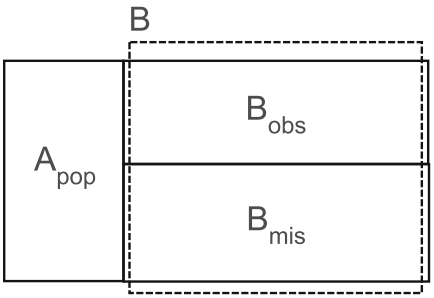

Fig. 1 Setting for fully synthetic data generation. A attributes whose values are available for all population subjects. B: attributes whose values are only available for the subjects in the original data set (B_{obs}: available values of B attributes; B_{mis}: missing values of B attributes)

5.1 Composability

Differential privacy offers strong composability properties. We focus on two of them [20]: sequential composition and parallel composition. Sequential composition refers to a sequence of computations, each of them providing differential privacy in isolation, providing also differential privacy in sequence.

Theorem 1 *Let $\kappa_i(D)$, for some $i \in I$, be computations over* D *providing ε_i-differential privacy. The sequence of computations $(\kappa_i(D))_{i\in I}$ provides $(\sum_{i\in I} \varepsilon_i)$-differential privacy.*

Parallel composition refers to several ε-differentially private computations each on data from a disjoint set of subjects yielding ε-differentially private output on the data from the pooled set of subjects.

Theorem 2 *Let $\kappa_i(D_i)$, for some $i \in I$, be computations over D_i providing ε-differential privacy. If each D_i contains data on a set of subjects disjoint from the sets of subjects of D_j for all $j \neq i$, then $(\kappa_i(D_i))_{i\in I}$ provides ε-differential privacy.*

The above composition properties hold for differentially private query answers and also for differentially private data sets. For the case of data sets, sequential composition can be rephrased as: the release of ε_i-differentially private data sets D_i, for some $i \in I$, is $(\sum_{i\in I} \varepsilon_i)$-differentially private. Sequential composition says that by accumulating differentially private data about a set of individuals, differential privacy is not broken, but the level of privacy decreases. Parallel composition can be rephrased as: The release of ε-differentially private data sets D_i referring to disjoint sets of individuals, for some $i \in I$, is ε-differentially private.

5.2 Computational Cost

Computing the value of a function f in a differentially private manner usually boils down to computing the value of f and then adding some noise to it [11,22]. The amount of noise required depends on the sensitivity of f (how much changes in a single record alter the value of f). Computing the sensitivity for an arbitrary function f can be a complex task. In such

cases, the sample and aggregate framework [22,26], based on computing the value of f on several population samples and then aggregating, is a good option. However, we are mainly interested in the computational cost associated with the generation of a differentially private data set.

For the case of a synthetically generated differentially private data set, the computational cost is highly dependent on the model used to approximate the distribution of the population. Hence, we restrict ourselves to the alternative approach that consists of computing a differentially private histogram. The cost of computing this histogram is proportional to the number of histogram bins. When the number of attributes is small, we can afford defining a set of prefixed bins for each attribute and generating the bins for the joint histogram by taking Cartesian products. However, since the number of bins for the joint histogram grows exponentially with the number of attributes, this approach is only applicable to data sets with a small number of attributes. Several approaches have been proposed to overcome this issue: In [5], bins are adjusted to the underlying data; in [32], the dependency between attributes is analyzed to limit the number of attributes in the histograms.

For the case of aggregation plus noise, the computational cost and the amount of noise needed to attain differential privacy are determined by the type of aggregation used. In [29], a multivariate microaggregation function is used to generate clusters containing k records and the average record of each cluster is computed. In [25], individual ranking microaggregation is used to group records and the average record is computed for each attribute. The cost of both proposals is quasilinear: $\mathcal{O}(n \ln n)$ on the number of records n of the data set.

5.3 Linkability

Like we noted above, when generating a differentially private synthetic data set, we can generate either a partially synthetic data set or a fully synthetic data set. Strictly speaking, a partially synthetic data set generated from an ε-differentially private model (histogram) of the original data does not satisfy ε-differential privacy. However, we consider partially synthetic data sets here for two main reasons. The first one is that the protection they offer may be enough in some situations. The second reason is that partial synthesis usually allows a very accurate linkage of records.

If the values used for the linkage between synthetic data sets have not been synthesized, we can expect perfectly accurate linkages. For a partially synthetic data set, these values would usually correspond to all the values of non-sensitive attributes and to non-sensitive values of sensitive attributes. For fully synthetic data sets, these values correspond to the attributes that are known for the entire population (the ones that correspond to matrix A in Fig. 1). However, there is a difference between partially and fully synthetic data that can have a great impact on linkability. In partially synthetic data, the subjects in the original data set are the ones that are present in the synthetic data set; thus, if data about a subject are collected and anonymized independently, we will be able to link the records. In fully synthetic data, a new sample from the underlying population is used to generate the new data set; thus, even if data about the same subject are collected by different sources, there is no guarantee that the subject will be present in the synthetic data sets.

For the case of a data set generated by masking the values of the original records, the discussion is similar. If the values used in the linkage have not been masked, then we can expect perfect linkage accuracy. However, it must be kept in mind that the values of the masked attribute are no longer the original values.

6 Conclusions

This paper has examined the privacy concerns raised by big data. In particular, big data seem to clash with preexisting private data management principles such as consent, purpose limitation, necessity and data minimization, transparency and openness, and individual rights to access, rectify and erase.

Although data anonymization is not an all-encompassing solution for privacy in big data (for example, it may thwart some types of data analysis), it can certainly be a useful tool to deal with the above clashes. However, the particular characteristics of big data (mainly the linkability requirement and the accumulation of data collected from multiple sources) challenge usual SDC approaches.

Privacy models are a relatively new development in the computer science literature on disclosure risk limitation. We have evaluated k-anonymity and ε-differential privacy in terms of composability, computational cost and linkability, all of them essential properties when dealing with big data. The main limitation of k-anonymity is related to composability: The release of several k-anonymous data sets may lead to re-identification of individuals. In contrast, differential privacy has strong composition properties: Releasing multiple differentially private data sets may, of course, increase the risk of disclosure, but one still has differential privacy (possibly with a greater value of the parameter and hence less privacy). Regarding computational cost and linkability, the exact performance of these privacy models depends on the approach used to generate the anonymized data, but there are options to deal with them.

In addition to adapting current privacy models for operation with big data, future research avenues include coming up with new privacy models designed from scratch with big data requirements in mind. In fact, changes may even be necessary in the definition of privacy guarantees, in order for

these to be naturally compatible with data continuously and massively collected from multiple data sources.

Acknowledgments The following funding sources are gratefully acknowledged: Government of Catalonia (ICREA Acadèmia Prize to the second author and Grant 2014 SGR 537), Spanish Government (projects TIN2011-27076-C03-01 "CO-PRIVACY" and TIN2014-57364-C2-1-R "SmartGlacis"), and European Commission (Project H2020-644024 "CLARUS"). The authors are with the UNESCO Chair in Data Privacy. The views in this paper are the authors' own and do not necessarily reflect the views of UNESCO.

References

1. Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, Zhu A (2005) Anonymizing tables. In: Eiter T, Libkin L (eds) Database theory—ICDT 2005, vol 3363., Lecture Notes in Computer Science. Springer, Berlin, p 246–258
2. Barbaro M, Zeller T (2006) A face is exposed for AOL searcher no. 4417749. New York Times, August 14
3. Brookman J, Hans GS (2013) Why collection matters: surveillance as a de facto privacy harm. In: Big data and privacy: making ends meet. The center for internet and society - Stanford Law School
4. Chen A (2010) Gcreep: google engineer stalked teens, spied on chats. Gawker, New York
5. Cormode G, Procopiuc C, Srivastava D, Shen E, Yu T (2012) Differentially private spatial decompositions. In: Proceedings of the 2012 IEEE 28th international conference on data engineering. ICDE'12, Washington, DC, USA. IEEE Computer Society, p 20–31
6. Danezis G, Domingo-Ferrer J, Hansen M, Hoepman J-H, Le Métayer D, Tirtea R, Schiffner S (2015) Privacy and data protection by design—from policy to engineering. Technical report, ENISA
7. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Comm ACM 51(1):107–113
8. Domingo-Ferrer J, Torra V (2005) Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min Knowl Discov 11(2):195–212
9. Duhigg C (2012) How companies learn your secrets. New York Times Magazine, February 16
10. Dwork C (2006) Differential privacy. In: Bugliesi M, Preneel B, Sassone V, Wegener I (eds) Automata, languages and programming, vol 4052., Lecture notes in computer science. Berlin, Springer, p 1–12
11. Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T (eds) Proceedings of the third conference on the theory of cryptography, vol 3876., lecture notes in computer science. Springer, p 265–284
12. Ganta SR, Kasiviswanathan SP, Smith A (2008) Composition attacks and auxiliary information in data privacy. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'08, New York, NY, USA. ACM, p 265–273
13. Hansell S (2006) AOL removes search data on vast group of web users. New York Times, August 8
14. Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES, Spicer K, de Wolf P-P (2012) Statistical disclosure control. Wiley, New York
15. LeFevre K, DeWitt DJ, Ramakrishnan R (2005) Incognito: efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD international conference on management of data, SIGMOD'05, New York, NY, USA. ACM, p 49–60
16. LeFevre K, DeWitt DJ, Ramakrishnan R (2006) Mondrian multidimensional k-anonymity. In: Proceedings of the 22nd international conference on data engineering, ICDE'06, Washington, DC, USA. IEEE Computer Society
17. Li N, Li T, Venkatasubramanian S (2007) t-Closeness: privacy beyond k-anonymity and l-diversity. In: Chirkova R, Dogac A, Özsu MT, Sellis TK (eds) Proceedings of the 23rd IEEE international conference on data engineering (ICDE 2007), p 106–115
18. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007) l-diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov Data, 1(1):3
19. Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L (2008) Privacy: theory meets practice on the map. In: Proceedings of the 2008 IEEE 24th international conference on data engineering, ICDE'08, Washington, DC, USA. IEEE Computer Society, p 277–286
20. McSherry FD (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. ACM, p 19–30
21. Meyerson A, Williams R (2004) On the complexity of optimal k-anonymity. In: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, PODS'04, New York, NY, USA. ACM, p 223–228
22. Nissim K, Raskhodnikova S, Smith A (2007) Smooth sensitivity and sampling in private data analysis. In: Proceedings of the thirty-ninth annual ACM symposium on the theory of computing, STOC'07, New York, NY, USA. ACM, p 75–84
23. Oganian A, Domingo-Ferrer J (2001) On the complexity of optimal microaggregation for statistical disclosure control. Stat J UN Econ Comm Eur 18:345–354
24. Samarati P (2001) Protecting respondents' identities in microdata release. IEEE Trans Knowl Data Eng 13(6):1010–1027
25. Sánchez D, Domingo-Ferrer J, Martínez S (2014) Improving the utility of differential privacy via univariate microaggregation. In: Domingo-Ferrer J (ed) Privacy in statistical databases, vol 8744, lecture notes in computer science. Springer, New York, pp 130–142
26. Smith A (2011) Privacy-preserving statistical estimation with optimal convergence rates. In: Proceedings of the forty-third annual ACM symposium on theory of computing, STOC'11, New York, NY, USA. ACM, p 813–822
27. Solove DJ (2011) Nothing to hide: the false tradeoff between privacy an security. Yale University Press, New Haven
28. Soria-Comas J, 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), p 1–8
29. Soria-Comas J, Domingo-Ferrer J, Sánchez D, Martínez S (2014) Enhancing data utility in differential privacy via microaggregation-based k-anonymity. VLDB J 23(5):771–794
30. Xiao Y, Xiong L, Yuan C (2010) Differentially private data release through multidimensional partitioning. In: Proceedings of the 7th VLDB conference on secure data management, SDM'10. Springer, Berlin, p 150–168
31. Xu J, Zhang Z, Xiao X, Yang Y, Yu G (2012) Differentially private histogram publication. In: Proceedings of the 2012 IEEE 28th international conference on data engineering, ICDE'12, Washington, DC, USA. IEEE Computer Society, p 32–43
32. Zhang J, Cormode G, Procopiuc CM, Srivastava D, 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. ACM, p 1423–1434