Anonymization in the Time of Big Data

Josep Domingo-Ferrer and Jordi Soria-Comas

Universitat Rovira i Virgili
Dept. of Computer Engineering and Mathematics
UNESCO Chair in Data Privacy
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
{josep.domingo,jordi.soria}@urv.cat

Abstract. In this work we explore how viable is anonymization to prevent disclosure in structured big data. For the sake of concreteness, we focus on k-anonymity, which is the best-known privacy model based on anonymization. We identify two main challenges to use k-anonymity in big data. First, confidential attributes can also be quasi-identifier attributes, which increases the number of quasi-identifier attributes and may lead to a large information loss to attain k-anonymity. Second, in big data there is an unlimited number of data controllers, who may publish independent k-anonymous releases on overlapping populations of subjects; the k-anonymity guarantee does not longer hold if an observer pools such independent releases. We propose solutions to deal with the above two challenges. Our conclusion is that, with the proposed adjustments, k-anonymity is still useful in a context of big data.
Keywords: Data anonymization, big data, k-anonymity, curse of dimensionality, multiple releases.

1 Introduction

When releasing microdata files for secondary use, the privacy of the subjects to whom the data refer must be taken into consideration. Statistical disclosure control (SDC) addresses this by altering the original data with the aim of protecting the privacy of the subjects while preserving sufficient utility for statistical analyses. In other words, a data user should not be able to accurately infer anything about a specific subject, but she should be able to accurately infer properties of subgroups of subjects. The SDC literature is quite rich; see [10, 20, 6] for comprehensive surveys.

Among the SDC methods used, anonymization is of particular relevance. The goal of anonymization of a microdata set is to blur the relation between each record and the corresponding subject. The usual anonymization approach is to classify the attributes into two categories: quasi-identifier attributes (QI) and confidential attributes. Confidential attributes are those that contain the sensitive information whose disclosure we want to prevent. An attribute is classified as a quasi-identifier attribute if it can help re-identify the subject corresponding to a record (link the record to a specific subject). The usual quasi-identifier attributes are socio-demographic attributes such as age, sex, gender, ZIP code, etc. More strictly, any attribute that is known to be externally available in association with a specific subject should be classified as a quasi-identifier attribute. The idea underlying anonymization is to alter the values of quasi-identifier attributes in the released data set in order to prevent linkage of its records to subjects in external databases containing identifiers and some of the quasi-identifier attributes; the ultimate goal is to prevent observers from learning the confidential attribute values of identified subjects.

With the development of information and communication technologies, data collection capabilities have witnessed a dramatic growth. The large amount of data available, together with the use of innovative data analysis techniques (whose goal is to generate knowledge rather than test hypothesis), have led to the development of big data. Previously, data collection (for example, by national statistical institutes) was a planned process that culminated in a structured data set. In a time when data were costly to acquire, the planning of data collection aimed at minimizing the resources required to collect the data, while ensuring the latter would be valid to make inferences on the population. In essence, we had scarce but accurate data. With big data, the situation is the opposite. Data are not scarce anymore; rather, they effortlessly and routinely flow in from different (and apparently unrelated) sources. There is no need for planning expensive data gathering processes: by living their normal lives, subjects generate lots of data (through queries submitted to a web search engine, emails, social media activities, web browsing, spatiotemporal traces of mobile devices, etc.). Big data may not be as accurate as traditional data, and they may be more complex to analyze (they may require automated analysis of textual information, merging of data coming from different sources, etc.). However, the sheer volume of information available makes up for the lost accuracy, and the very diverse nature of the data can make them suitable for very diverse analyses.

In a big data context, important assumptions previously made in data anonymization procedures seem unrealistic. The determination of the quasi-identifier attributes was already controversial in the past because an observer with privileged knowledge of a target subject could use attributes not publicly known to most people as quasi-identifier attributes to re-identify the target's record. Despite this possibility, the fact that there were relatively few data controllers, that these were trusted, and that most of them collected the data under a strong pledge of privacy made it reasonable to classify attributes into two disjoint sets, the quasi-identifier attributes and the confidential attributes. With the development of big data, the number of data controllers has grown dramatically and it is unrealistic to view all of them as trusted. In fact, an untrusted controller can unduly leak confidential attributes, so that they can be used as quasi-identifier attributes by whoever acquires them. This new scenario must be taken into account when using SDC to protect a data set before release; otherwise the privacy protection may be insufficient. In particular, in the case of anonymization, we must be aware that any attribute is a potential quasi-identifier attribute.

Even though big data may seem to undermine the foundations of data anonymization by turning each attribute into a potential quasi-identifier attribute, anonymization remains an important desideratum in day-to-day practice. It is not only that its intuitiveness makes it a very appealing SDC method but also that many data protection regulations [2, 3, 1] are built around anonymization. Since this situation is unlikely to change, we should strive to carry out anonymization in a safe manner.

Contribution and plan of this paper

In this work, we analyze the impact of big data on current anonymization practices, most notably on the best-known anonymization-based privacy model, namely k-anonymity. We propose alternatives to maintain the disclosure control guarantees of k-anonymity in the big data world. Our aim is not so much to propose novel anonymization methods that are resilient to observers in a big data environment, but to find ways to adapt the current k-anonymity algorithms so that they can better deal with the new state of things. In other words, we do not seek to propose k-anonymity methods capable of dealing with all big data formats (e.g. textual data, video, audio, etc.) but rather enhance existing methods for structured data so that they remain effective in a big data environment.

The remainder of this paper is structured as follows. Section 2 gives background on anonymization concepts, data protection regulations, k-anonymity and some shortcomings of k-anonymity. Section 3 analyzes the two main challenges encountered by k-anonymity in a big data setting: first, confidential attributes are likely to be also QI attributes, which may lead to a large number of quasi-identifiers and hence to unacceptable utility loss; second, different data controllers may disseminate multiple independent k-anonymous releases on overlapping populations of subjects, which compromises the k-anonymity guarantee. This section does not only identify the two aforementioned challenges, but it also proposes possible solutions. Finally, conclusions and future research issues are gathered in Section 4.

2 Background

2.1 Anonymization

Anonymization is the process of masking the correspondence between records in a data set and the subjects to whom they correspond. It entails two main steps:

- *De-identification.* This consists of suppressing the identifier attributes (those that unambiguously identify a subject). Example identifier attributes include: name, social security number, email address, etc.
- *Quasi-identifier masking.* Quasi-identifier attributes can also be used to reidentify the subject to whom a record corresponds. However, unlike identifiers, quasi-identifier attributes should not be removed in general, because of the analytical value they convey. In fact, most attributes have the potential of being quasi-identifier attributes, depending on the observer's knowledge. Hence, masking them is much better than removing them.

2.2 Data protection regulations and anonymization

The general view of data protection regulations is that anonymized data are not personal (re-identifiable) anymore. Two remarkable examples of such regulations are:

- U.S. Health Insurance Portability and Accountability Act (HIPAA) [1]. It was enacted in 1996 by the U.S. Congress in an attempt to improve healthcare and health insurance. According to Title II of the HIPAA, health information is not re-identifiable if either (i) a person with appropriate knowledge justifies that the risk is very small, or (ii) several attributes are removed or generalized to a specified level (e.g. names, telephone numbers, and social security numbers are removed; geographic units and dates are generalized).
- E.U. General Data Protection Regulation (GDPR) [3]. This is the European Union's forthcoming data protection regulation. In comparison with the current Directive 95/46/EC [2], that it will replace, the GDPR is more ambitious in terms of uniformity across the E.U., enforcement mechanisms and fines. The GDPR targets any personal (re-identifiable) data whatever the nature of the data is. Thus, non-re-identifiable (e.g. anonymized) data are not under the umbrella of the GDPR. However, the GDPR does not specify when data are (or are not) re-identifiable.

2.3 *k*-Anonymity

k-Anonymity [16, 15] is a well-known data privacy model that is based on anonymization. k-Anonymity seeks to guarantee at least a certain level of anonymity for the records in a data set. To that end, k-anonymity assumes that the attributes an observer can use to re-identify a record can be determined beforehand by the data controller, and makes each record indistinguishable with regard to these attributes within a group of k records.

In k-anonymity, the term *quasi-identifier* refers to a *set of attributes* whose combined values can lead to record re-identification. Thus, it focuses on controlling the release of such attributes. This is similar to the above-mentioned notion of quasi-identifier attribute but it is not equivalent. The attributes that form a quasi-identifier are quasi-identifier attributes, but a quasi-identifier attribute alone is not necessarily a quasi-identifier.

Definition 1 (Quasi-identifier). *Let X be a data set with attributes $X_1, \ldots, X_m$. A quasi-identifier of X is a subset of attributes $\{X_{i_1}, \ldots, X_{i_j}\} \subset \{X_1, \ldots, X_m\}$ whose combined values can lead to re-identification. Denote by QI_X the set of quasi-identifiers of X.*

Whether a set of attributes should be considered a quasi-identifier depends on the external information available to observers. The set of quasi-identifiers usually includes all the combinations of attributes that the data releaser reasonably assumes to be publicly available in external non-de-identified data sets (electoral rolls, phone books, etc.).

To prevent re-identification of records based on a quasi-identifier, k-anonymity requires each combination of values of the quasi-identifier attributes to be shared by k or more records.

Definition 2 (k-Anonymity). *Let X be a data set and let QI be the maximal quasi-identifier in QI_X, that is, the one containing all quasi-identifier attributes. The data set X is said to be k-anonymous if each combination of values of the attributes in QI that appears in X is shared by k or more records.*

In a k-anonymous data set, each group of k or more records sharing the values of all quasi-identifier attributes is called a *k-anonymous class*.

Several algorithms to generate k-anonymous data sets have been proposed. The most common ones are based on generalization and suppression [11, 16] and on microaggregation [7].

2.4 Shortcomings of k-anonymity

The kind of disclosure risk limitation offered by k-anonymity is very intuitive: if we prevent records from being linked to the corresponding subjects, releasing the data set will not result in any disclosure. Indeed, this view of disclosure risk limitation is quite effective when the observer cannot decide by looking at the k-anonymous classes if a given subject is in the data set or not. However, the usual observer considered in k-anonymity knows that his target subject is in the k-anonymous data set. Under this observer model, the target subject can be linked to a k-anonymous class.

When all records in a k-anonymous class have similar values for the confidential attributes, an observer linking a subject to that class can infer confidential information on that subject. In other words, while k-anonymity provides protection against identity disclosure (at most a subject can be linked to her record's class), it cannot prevent attribute disclosure when the values of the confidential attributes are similar within a k-anonymous class. Two simple attacks have been proposed in the literature that exploit the lack of variability in the confidential attributes: the homogeneity attack and the background knowledge attack [13]. In the former, all the records in a k-anonymous class share the same value for the confidential attributes; thus, k-anonymity fails to provide any protection. In the background knowledge attack, the variability of the confidential attributes within the class is small and the observer possesses some background information that allows him to further restrict the feasible values of the confidential attributes for the target subject. To achieve protection against attribute disclosure, several refinements of the basic notion of k-anonymity have been proposed (e.g. l-diversity [13], t-closeness [12], β-likeness [5], etc.). All these require at least a certain level of variability in the confidential attributes within each k-anonymous class. Although attribute disclosure is an important issue in k-anonymity, it is not specifically related to the big data scenario and it has already been extensively treated in the literature (for example, in [8], in addition to the previous references).

How to safely release multiple k-anonymous data sets sharing some subjects and some confidential attributes is an open issue [19, 14]. An example would be when two hospitals with possibly some common patients release respective k-anonymous patient discharge data sets including confidential attributes such as diagnosis and/or measurements on the patients. In this situation, the k-anonymity disclosure limitation guarantee is preserved over the two pooled releases only if any two k-anonymous classes coming from different k-anonymous releases have either all subjects in common or no subject in common. If two k-anonymous classes have k' subjects in common, with with $0 < k' < k$, then only k'-anonymity is provided for those common subjects and only $(k - k')$-anonymity for the non-common subjects (note that comparing the confidential attribute values is likely to help determine the common subjects).

Hence, preserving k-anonymity in case of multiple releases requires a coordination that is easiest if all releases are performed by the same data controller and they all relate to the same population of subjects. When the data controller is the same but the populations of subjects in the releases are neither disjoint nor identical, the problem is more complex but still manageable. In particular, [14] proposes an approach to release k-anonymous incremental updates, which may modify the population by adding or removing some subjects, based on coarsening subsequent data releases given the previous ones to ensure that k-anonymity is still satisfied. Since the focus of our work here is big data, we are more interested in k-anonymous data releases independently performed by different data controllers.

3 k-Anonymity in a big data context

We leave aside the difficulties of k-anonymity with low-variability confidential attributes and with multiple releases reviewed in the previous section, and we move to challenges inherent to big data. In a big data setting, there is a very large number of data controllers, so the most likely option in case of multiple releases on non-disjoint populations is that they are performed by different controllers. Let us explore the ramifications of this situation.

3.1 Confidential attributes as QI attributes

Like most privacy models, k-anonymity makes some assumptions on the side knowledge available to the observer. The disclosure risk limitation procedure is then tailored to reach the privacy guarantee stated by the model against such an observer. If the assumptions were too optimistic (that is, the actual observer has more side knowledge than assumed), the intended privacy guarantee may not be attained.

A key assumption in k-anonymity is that there is a clear distinction between quasi-identifiers and confidential attributes. In a big data setting, we argue that such a distinction is no longer tenable. There is a large number of data controllers, a lot of which collect data unawares of the subjects (by inspecting emails, tracking web navigation, recording web searches, recording geospatial information available from mobile devices, etc.). Hence, it is unrealistic to think that all controllers can be trusted. In turn, untrusted controllers may share, leak or sell any confidential attribute, which can then be used by the party acquiring it as a quasi-identifier attribute to improve re-identification attacks. Even if an untrusted controller does not share confidential attributes, he can use them himself to mount re-identification attacks against k-anonymous data sets released by other controllers.

Another usual assumption in k-anonymity is that there is a single confidential attribute. This simplification is innocuous when the confidential data cannot be used to improve record re-identification. However, when they *can* be used, as we argue is the case for big data, this simplification may mask the problem: if there is a single confidential attribute, an observer using it to improve record re-identification does not learn much, because he already knew the confidential data in the first place. To avoid this kind of (wrong) justification for excluding confidential attributes from quasi-identifiers, we consider a data set with multiple confidential attributes. Hence, our data set definition is as follows.

Definition 3. *The original data set is a table $T(A_1, \dots, A_r, C_1, \dots, C_s)$ where $A_1, \dots, A_r$ are quasi-identifier attributes and $C_1, \dots, C_s$ are confidential and quasi-identifier attributes at the same time.*

The usual approach in k-anonymity is to mask the QI attributes in such a way that each subject can be linked to k records with the same probability. For instance, when using generalization, we look for a minimal generalization of the QI that satisfies the k-anonymity requirement. Because the confidential attributes are also QI, all the attributes must be generalized. This naive approach has an important drawback: the curse of dimensionality, which states that, as the number of QI attributes grows, all the discriminatory information in the data is lost in order to achieve k-anonymity [4]. Thus, adding all confidential attributes to the set of QI attributes can be devastating for the data utility. To avoid increasing the number of QI attributes so much, we propose to release a separate k-anonymous data set for each confidential attribute. That is, rather than generating a k-anonymous $T_k(A_1, \dots, A_r, C_1, \dots, C_s)$, we generate separate k-anonymous versions of the s following data sets:

$$T_k^1(A_1, \dots, A_r, C_1), \dots, T_k^s(A_1, \dots, A_r, C_s). \qquad (1)$$

It might seem that we can release the data sets given by Expression (1) without generalizing the confidential attribute in each data set. After all, if the observer uses the confidential attribute to re-identify a subject, then it means that he already knew the value of the confidential attribute and, therefore, he does not learn anything. The previous reasoning is a (also flawed) variant of the abovementioned justification to exclude confidential attributes from the set of QI attributes. It overlooks that the observer can leverage the knowledge of the confidential attribute value of a subject to improve the re-identification of the other subjects in the same k-anonymous class: this amounts to downgrading k-anonymity to $(k-1)$-anonymity. Therefore, when generating the k-anonymous versions of $T_k^1, \ldots, T_k^s$, the confidential attribute must also be generalized.

On the other hand, in Section 2.4 we said that, when releasing multiple k-anonymous data sets corresponding to original data sets sharing some subjects *and some confidential attributes*, two k-anonymous classes coming from two different releases should either have all subjects in common or no subject in common. The reason was that, if two classes had $0 < k' < k$ subjects in common, only k'-anonymity was guaranteed for the common subjects and only $(k-k')$-anonymity was guaranteed for the non-common subjects. When stating so, there was an implicit assumption that the records corresponding to the k' common subjects could be determined within each class, by linking them via the confidential attributes. However, in the k-anonymous releases of the datasets in Expression (1), there are *no common* confidential attributes among releases. If, beyond being different, the confidential attributes C_i and C_j in two releases are uncorrelated, then common subjects cannot be determined. Therefore, in case of uncorrelated confidential attributes, the data sets in Expression (1) can be k-anonymized independently, using standard k-anonymization algorithms in the literature.

The situation changes if there is a strong correlation among the confidential attributes in the different releases. In this case, the partition of subjects into k-anonymous classes in T_k^i and T_k^j must be the same, for all $i \neq j$. This constraint limits our ability to use existing algorithms out of the box: the quasi-identifier is not the same in both tables, because the confidential attributes are also QI attributes and they are different. Thus, using an existing k-anonymization algorithm independently on both tables is likely to yield different k-anonymous classes. To overcome this issue, we propose to perform the generalization in two steps: the first step generates k-anonymous classes taking into account only attributes $A_1, \ldots, A_r$; the second step generalizes the confidential attribute for it to have a single value within each class generated in the previous step. Algorithm 1 formalizes this two-step k-anonymization.

To guard against attribute disclosure, it is relatively simple to adjust Algorithm 1 to ensure a certain level of variability of the confidential attributes $C_1, \ldots, C_s$ within each k-anonymous class. This minimum level of variability must be taken into account when generating the k-anonymous classes G_k.

Even though Algorithm 1 is not needed to enforce k-anonymity when the confidential attributes are uncorrelated, there is an advantage in using it always: we can merge the k-anonymous data sets $T_k^1, \ldots, T_k^s$ into a single k-anonymous data set that contains all the confidential attributes.

3.2 Independent k-anonymous data releases on non-disjoint populations

k-Anonymity was designed to limit the disclosure risk of a single data release. In general, if several k-anonymous releases are pooled by the observer, there is no guarantee that k-anonymity holds anymore and attribute disclosure can

Algorithm 1 Masking of an original data set into several k-anonymous data sets having the same subject partition and the same non-confidential quasi-identifier attributes $A_1, \ldots, A_r$ and having each a single non-common confidential quasi-identifier attribute.

```
let $T(A_1, \ldots, A_r, C_1, \ldots, C_s)$ be a table where $C_1, \ldots, C_s$ are confidential attributes.

let $G_k$ be the set of $k$-anonymous classes generated for the attributes $A_1, \ldots, A_r$
for $i = 1$ to $s$
    $T_k^i(A_1, \ldots, A_r, C_i)$ :=$G_k$ + generalization of $C_i$ values within each $k$-anonymous class
end for

return $T_k^1, \ldots, T_k^s$
```

occur. A paradigmatic example consists of two k-anonymous data releases in which the confidential attributes are not considered to be QI attributes (and hence stay unmodified) and there are two k-anonymous classes that differ in one subject. By comparing the values of the confidential attributes in these k-anonymous classes, an observer can learn the value of the confidential attributes of that subject (attribute disclosure). Furthermore, by comparing the way the QI attribute values for that subject have been modified in the two k-anonymous releases, the observer may gain some information on the original values of those QI attributes, which may facilitate re-identifying that subject with probability greater than $1/k$ (re-identification disclosure).

As said in Section 2.4, if all the data releases are performed by the same data controller, the above attack can be avoided by either making sure that the same k-anonymous class of records are used when the population does not change between data releases or by adjusting the classes of records to satisfy k-anonymity when the population changes. The thorniest scenario is when independent k-anonymous data releases are performed by different data controllers, which is precisely the most common case in big data settings. We tackle this scenario next.

The example given at the beginning of this section shows that controlling the disclosure risk across multiple and independent k-anonymous data releases is not feasible in general. Even if we follow the approach suggested in Section 3.1 for the case of confidential attributes that are also QI attributes (thus making sure that every subject can be linked to k records), *the change in the generalization of the confidential attribute between k-anonymous classes of records that differ in one subject may reveal the actual confidential attribute value of that subject.* In particular, for the case of a continuous confidential attribute that has been generalized as an interval encompassing all values of the confidential attribute in the k-anonymous class, the confidential attribute may be disclosed for the subjects that have the largest or the smallest confidential attribute value in the class.

To minimize the risk of disclosing confidential information, we propose to replace the generalization of the confidential attribute by an alternative aggregation operation that is less sensitive to changes in the components of the k-anonymous class. We list below two possible alternatives:

- Mean of the values of the confidential attribute. By using the mean, we make sure that each subject contributes to the aggregation by a proportion of $1/k$. Despite this fact, the mean can still be sensitive to the presence of subjects with very extreme values.
- Median of the values of the confidential attribute. In comparison to the mean, the median is less sensitive to the extreme values. However, because the median reflects the value of a specific subject, a change in the median between k-anonymous classes of records that differ in one subject may reveal the exact value of the confidential attribute for that subject.

Apart from the above-mentioned respective shortcomings of the mean and the median, the fact that both are deterministically computed from the values in the k-anonymous classes introduces another problem, described next. In a big data scenario and given a target subject, it is not unrealistic to assume that the values of the confidential attribute are available to the observer for all the subjects other than the target subject in latter's k-anonymous class. In this case, when using the mean, the observer can determine the target subject's confidential attribute value; when using the median, the observer can determine it in some cases.

To prevent aggregation-related disclosures, we should insert some uncertainty in the aggregation. Differential privacy [9] can be used to introduce such an uncertainty in the case of the mean. Differential privacy aims at making the result of a computation similar between data sets that differ in one subject (neighbor data sets). By adding noise to the result of the computation, differential privacy makes sure that the probability of getting a specific result is similar between neighbor data sets. Usually, one adds Laplace-distributed noise with 0 mean and scale parameter that depends on the maximum change in the computation between neighbor data sets (a.k.a. the global sensitivity). For the case of adding differentially private noise to the mean of k-anonymous classes, the global sensitivity is the size of the domain of the confidential attribute divided by k [17, 18].

4 Conclusions and future work

This work has tried to answer the question of whether anonymization, and more precisely k-anonymity, is still a valid approach to prevent disclosure in the current big data situation. We have analyzed from a qualitative perspective how k-anonymity can be adapted for statistical disclosure control of big data. In particular, we have addressed the following two challenges and we have proposed solutions for them:

- In big data, any attribute, including the confidential attributes, can be a QI attribute. However, reaching k-anonymity with so many QI attributes in a straightforward way is subject to the so-called curse of dimensionality: a lot of information is lost. We have given a k-anonymization algorithm that circumvents this dimensionality problem.
- In big data there is a great number of data controllers and these are likely to generate multiple independent k-anonymous releases on overlapping populations of subjects. Keeping the disclosure risk at the level that is guaranteed for a single k-anonymous data release is not feasible: for example, an observer can compare k-anonymous data sets that differ in one subject and discover the confidential attribute values of that subject. To minimize the risk of disclosure, we have proposed to replace the usual generalization performed within each k-anonymous class by an alternative aggregation operation that is less sensitive to changes in the subjects of the class. In particular, we have explored using the mean and the median. The fact that these are computed deterministically from the confidential values of the subjects in the k-anonymous class is also a potential source of disclosure. For the case of the mean, we have proposed to neutralize such disclosure using a ε-differentially private mean.

Our qualitative analysis seems to indicate that anonymization and k-anonymity are still useful in our big data world. As future work, we expect to come up with a more quantitative analysis, where we will evaluate the utility that can be achieved for a given anonymity guarantee. We will also explore how to randomize other aggregation operators beyond the mean for the case of multiple independent k-anonymous releases.

Acknowledgments and disclaimer

Partial support to this work has been received from the European Commission (projects H2020-644024 “CLARUS” and H2020-700540 “CANVAS”), from the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2014 SGR 537), and from the Spanish Government (project TIN2014-57364-C2-1-R “SmartGlacis” and TIN 2015-70054-REDC). The authors are with the UNESCO Chair in Data Privacy, but the views in this paper are the authors' own and are not necessarily shared by UNESCO.

References

1. U.S. Health Insurance Portability and Accountability Act (HIPAA, Pub. L 104-191, 110 Stat. 1936). Aug. 21, 1996.
2. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Communities*, pp. 31–50, Nov. 23, 1995.

3. European Parliament legislative resolution of 14 April 2016 on the Council position at first reading with a view to the adoption of a regulation of the European Parliament and of the Council 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 (General Data Protection Regulation). Apr. 2016.
4. C.C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases-VLDB 2005*, pp. 901–909, 2005.
5. J. Cao and P. Karras. Publishing microdata with a robust privacy guarantee. *Proceedings of the VLDB Endowment*, 5(11):1388–1399, 2012.
6. J. Domingo-Ferrer, D. Sanchez, and J. Soria-Comas. *Database anonymization: privacy models, data utility, and microaggregation-based inter-model connections*. Morgan & Claypool, 2016.
7. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Minining and Knowledge Discovery*, 11(2):195–212, 2005.
8. J. Domingo-Ferrer and V. Torra. A critique of k-anonymity and some of its enhancements. In *Proceedings of the 3rd International Conference on Availability, Reliability and Security-ARES 2008*, pp. 990–993. IEEE, 2008.
9. C. Dwork. Differential privacy. In *Automata, Languages and Programming*, LNCS 4052, pp. 1–12. Springer, 2006.
10. A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E.S. Nordholt, K. Spicer, and P.P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.
11. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering-ICDE'06*, pp. 25-37. IEEE, 2006.
12. N. Li, T. Li, and S. Venkatasubramanian. t-closeness: privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd International Conference on Data Engineering-ICDE'07*, pp. 106–115. IEEE, 2007.
13. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.
14. J. Pei, J. Xu, Z. Wang, W. Wang, and K. Wang. Maintaining k-anonymity against incremental updates. In *19th International Conference on Scientific and Statistical Database Management-SSBDM '07*. IEEE, 2007.
15. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.
16. P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.
17. D. Sánchez, J. Domingo-Ferrer, and S. Martínez. Improving the utility of differential privacy via univariate microaggregation. In *Privacy in Statistical Databases*, LNCS 8744, pp. 130–142. Springer, 2014.
18. D. Sánchez, J. Domingo-Ferrer, S. Martínez, and J. Soria-Comas. Utility-preserving differentially private data releases via individual ranking microaggregation. *Information Fusion*, 30:1 – 14, 2016.
19. K. Stokes and V. Torra. Multiple releases of k-anonymous data sets and k-anonymous relational databases. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 20(6):839–853, 2012.
20. L. Willenborg and T. DeWaal. *Elements of statistical disclosure control*. Springer-Verlag, New York, 2001.